Statistics Surveys
Vol. 9 (2015) 32–105
ISSN: 1935-7516
DOI: 10.1214/15-SS109

M-functionals of multivariate scatter

Lutz Dümbgen[*]

University of Bern, Sidlerstr. 5, CH-3012 Bern, Switzerland
e-mail: duembgen@stat.unibe.ch

Markus Pauly[†]

Ulm University, Helmholtzstr. 20, D-89081 Ulm, Germany
e-mail: markus.pauly@uni-ulm.de

and

Thomas Schweizer[*]

University of Bern, Sidlerstr. 5, CH-3012 Bern, Switzerland
e-mail: thomas-za.schweizer@ubs.com

Abstract: This survey provides a self-contained account of M-estimation of multivariate scatter. In particular, we present new proofs for existence of the underlying M-functionals and discuss their weak continuity and differentiability. This is done in a rather general framework with matrix-valued random variables. By doing so we reveal a connection between Tyler's (1987a) M-functional of scatter and the estimation of proportional covariance matrices. Moreover, this general framework allows us to treat a new class of scatter estimators, based on symmetrizations of arbitrary order. Finally these results are applied to M-estimation of multivariate location and scatter via multivariate t-distributions.

MSC 2010 subject classifications: 62G20, 62G35, 62H12, 62H99.
Keywords and phrases: Coercivity, convexity, matrix exponential function, multivariate t-distribution, scatter functionals, weak continuity, weak differentiablity.

Received January 2014.

Contents

[*]Supported by Swiss National Science Foundation (SNF).
[†]Supported by a fellowship within the postdoc programme of the German Academic Exchange Service (DAAD).

1. Introduction

The study of M-estimation for certain parameters or functionals of interest has a long history. Roughly speaking an M-estimator is the maximizer of a random criterion function depending on the data and corresponding to the estimation problem. Best known examples are maximum-likelihood estimators as well as robust estimators of location, e.g. the sample median, and scatter. In basic statistics courses it is shown that especially maximum-likelihood estimators are asymptotically normal and efficient under quite weak assumptions, see e.g. the graduate textbooks by Serfling (1980), Lehmann and Casella (1998) and van der Vaart (1998). Specific M-estimators of one- and multidimensional parameters can be shown to be asymptotically normal and quite efficient under even weaker assumptions, see e.g. Huber (1964; 1973), thus providing an interesting alternative to classical unbiased estimators.

In the present survey we consider M-estimates and functionals of multivariate location and scatter. Our purpose is to provide a concise but self-contained presentation of the main ideas and results in this context, the target audience

being researchers and advanced graduate students. The basic setting is as follows: Let P be a probability distribution on $\mathbb{R}^q$. Traditionally the center of P is defined to be the mean vector

$$\boldsymbol{\mu}(P) \;:=\; \int x\, P(dx),$$

assuming that $\int \|x\|\, P(dx) < \infty$. Assuming also that $\int \|x\|^2\, P(dx) < \infty$, the covariance matrix of P is defined as

$$\boldsymbol{\Sigma}(P) \;:=\; \int (x - \boldsymbol{\mu}(P))(x - \boldsymbol{\mu}(P))^\top\, P(dx),$$

where vectors are understood as column vectors and $(\cdot)^\top$ denotes transposition. Recall that for a random vector X with distribution P and any fixed vector $v \in \mathbb{R}^q$,

$$\mathbb{E}(v^\top X) \;=\; v^\top \boldsymbol{\mu}(P) \quad \text{and} \quad \mathrm{Var}(v^\top X) \;=\; v^\top \boldsymbol{\Sigma}(P) v.$$

Thus for a unit vector $v \in \mathbb{R}^q$, the spread of P in direction v may be quantified by $\sqrt{v^\top \boldsymbol{\Sigma}(P) v}$, the standard deviation of $v^\top X$.

There are various good reasons to use different definitions of the center $\boldsymbol{\mu}(P)$ and scatter matrix $\boldsymbol{\Sigma}(P)$ of the distribution P. For instance, suppose that P has a unimodal density f and is elliptically symmetric with center $\mu \in \mathbb{R}^q$ and symmetric, positive definite scatter matrix $\Sigma \in \mathbb{R}^{q \times q}$. That means, f may be written as

$$f(x) \;=\; \tilde{f}\big((x - \mu)^\top \Sigma^{-1} (x - \mu)\big)$$

for some decreasing function $\tilde{f} : [0, \infty) \to [0, \infty)$. Then it would be natural to define the center of P to be $\boldsymbol{\mu}(P) := \mu$, and a scatter matrix $\boldsymbol{\Sigma}(P)$ of P should be equal or at least proportional to Σ, even if $\int \|x\|\, P(dx)$ or $\int \|x\|^2\, P(dx)$ is infinite. A related issue is robustness: One would like $\boldsymbol{\mu}(P)$ and $\boldsymbol{\Sigma}(P)$ to change little if P is replaced with $(1 - \epsilon)P + \epsilon P'$ for some small number $\epsilon > 0$ and an arbitrary distribution P' on $\mathbb{R}^q$. Another way to define robustness is weak continuity: It would be desirable that $\boldsymbol{\mu}(P') \to \boldsymbol{\mu}(P)$ and $\boldsymbol{\Sigma}(P') \to \boldsymbol{\Sigma}(P)$ whenever $P' \to P$ weakly.

Some people may feel overwhelmed by the diversity of scatter functionals which are available. However, comparing two or more different scatter matrices $\boldsymbol{\Sigma}(P)$ allows one to find interesting structures in the distribution P. For an explanation of this paradigm and examples we refer to Nordhausen et al. (2008), Tyler et al. (2009) and the references cited therein.

A special class of location and scatter functionals are multivariate M-functionals. Introduced by Maronna (1976), their properties have been analyzed by numerous authors, an incomplete list of references being Huber (1981), Hampel et al. (1986), Tyler (1987a; 1987b), Kent and Tyler (1988; 1991) and Dudley et al. (2009). In particular, Dudley et al. (2009) prove existence and uniqueness of multivariate t-functionals of location and scatter, generalizing results of Kent and Tyler (1988; 1991). Moreover, they provide an in-depth analysis of weak continuity and differentiability of such functionals which implies consistency and

asymptotic normality of the corresponding estimators. Similar considerations have been made by Dümbgen (1998) for the special M-functional of scatter due to Tyler (1987a). As to the robustness of multivariate t-functionals of location and scatter in terms of so-called breakdown points, we refer to Dümbgen and Tyler (2005) and the references therein.

In many settings the location parameter $\boldsymbol{\mu}(P)$ is merely a nuisance parameter while the main interest lies on the scatter matrix $\boldsymbol{\Sigma}(P)$. Moreover, often one only needs to know $\boldsymbol{\Sigma}(P)$ up to a positive scaling factor, e.g. when defining principal components or correlations. On the other hand, a desirable feature is the following block independence property: If P describes the distribution of $X = [X_1^\top, X_2^\top]^\top$ with two stochastically independent random vectors $X_1 \in \mathbb{R}^{q(1)}, X_2 \in \mathbb{R}^{q(2)}$, then $\boldsymbol{\Sigma}(P)$ should be block diagonal, i.e.

$$\boldsymbol{\Sigma}(P) \;=\; \begin{bmatrix} \boldsymbol{\Sigma}_1(P) & 0 \\ 0 & \boldsymbol{\Sigma}_2(P) \end{bmatrix}$$

with $\boldsymbol{\Sigma}_i(P) \in \mathbb{R}^{q(i)\times q(i)}$. Unfortunately, the M-functionals just mentioned do not have this property. However, as explained later, any reasonable M-functional of scatter has the block independence property when it is applied to the symmetrized distribution $\mathcal{L}(X - X')$ with independent random vectors $X, X' \sim P$. (Here and throughout $\mathcal{L}(Y)$ denotes the distribution of a random variable Y, and $Y \sim Q$ is shorthand for "Y has distribution Q".) Note also that the symmetrized distribution $\mathcal{L}(X - X')$ is centered around $0 \in \mathbb{R}^q$, so we may avoid the estimation of a location parameter and focus on estimation of scatter only. This trick is used by many authors, e.g. Croux et al. (1994), Dümbgen (1998), Sirkiä et al. (2007), Nordhausen et al. (2008) and Tyler et al. (2009).

Applying the M-functionals $\boldsymbol{\mu}(\cdot)$ and $\boldsymbol{\Sigma}(\cdot)$ to the empirical distribution $\widehat{P}$ of independent random vectors $X_1, X_2, \ldots, X_n$ with distribution P yields *M-estimators* $\widehat{\mu} = \boldsymbol{\mu}(\widehat{P})$ and $\widehat{\Sigma} = \boldsymbol{\Sigma}(\widehat{P})$.

The remainder of this survey is organized as follows: In Section 2 we review the concepts of affine and linear equivariance and their main consequences. In Section 3 we motivate M-functionals of location and scatter by various maximum-likelihood and other estimation problems. After these introductory sections, we start with the main results about existence, uniqueness, weak continuity and differentiability of the M-functionals.

The main part of our paper is devoted to scatter-only functionals, treated in Sections 4, 5 and 6. This is done in a generalized framework with matrix-valued random variables. By doing so we reveal a connection between Tyler's (1987a) M-functional of scatter and the estimation of proportional covariance matrices as treated by Flury (1986), Eriksen (1987) and Jensen and Johansen (1987). Moreover, this general framework allows us to treat a new class of scatter estimators, based on symmetrizations of arbitrary order. Part of this material is new. Section 4 contains the main results about existence and uniqueness of the scatter functionals. Section 5 provides analytical tools to derive the aforementioned and later results. As realized by Auderset et al. (2005) in the context

of multivariate (real or complex) Cauchy distributions and by Wiesel (2012), among others, working with matrix exponentials and logarithms in a suitable way provides valuable new insights, and we are utilizing this approach, too. In particular, the target functions to be minimized turn out to be (strictly) convex in a certain sense which is essential for uniqueness. In our opinion, the resulting proofs are more intuitive than some derivations in the original papers. Based on the analytical results in Section 5, we discuss weak continuity and weak differentiability of scatter functionals in Section 6.

Finally, in Section 7 we review a trick by Kent and Tyler (1991) to treat location and scatter functionals based on multivariate t-distributions by means of the scatter-only methods. This allows one to prove weak differentiability and central limit theorems as in Dudley et al. (2009).

Various auxiliary results and most proofs are deferred to Section 8.

Notation Throughout this paper, the standard Euclidean norm of a vector $v \in \mathbb{R}^d$ is denoted by $\|v\| = \sqrt{v^\top v}$. For matrices $A, B \in \mathbb{R}^{q\times d}$ we use either the operator or the Frobenius norm,

$$\begin{aligned} \|A\| &:= \max_{v\in\mathbb{R}^d\setminus\{0\}} \frac{\|Av\|}{\|v\|} = \max_{v\in\mathbb{R}^d:\|v\|=1} \|Av\|, \\ \|A\|_F &:= \Big(\sum_{i,j} A_{ij}^2\Big)^{1/2} = \langle A, A\rangle^{1/2}, \end{aligned}$$

where

$$\langle A, B\rangle := \sum_{i,j} A_{ij}B_{ij} = \operatorname{tr}(A^\top B) = \operatorname{tr}(AB^\top).$$

Note that $\langle A, B\rangle$ defines an inner product on $\mathbb{R}^{q\times d}$. If $\operatorname{vec}(A)$ and $\operatorname{vec}(B)$ denote vectors in $\mathbb{R}^{qd}$ containing the columns of A and B, respectively, then $\langle A, B\rangle$ is just the usual inner product $\operatorname{vec}(A)^\top\operatorname{vec}(B)$. We shall consider the following subsets of $\mathbb{R}^{q\times q}$:

$$\begin{aligned} \mathbb{R}^{q\times q}_{\mathrm{ns}} &:= \big\{A \in \mathbb{R}^{q\times q} : A \text{ nonsingular}\big\}, \\ \mathbb{R}^{q\times q}_{\mathrm{sym}} &:= \big\{A \in \mathbb{R}^{q\times q} : A = A^\top\big\}, \\ \mathbb{R}^{q\times q}_{\mathrm{sym},\geq 0} &:= \big\{A \in \mathbb{R}^{q\times q}_{\mathrm{sym}} : A \text{ positive semidefinite}\big\} \\ &= \big\{A \in \mathbb{R}^{q\times q}_{\mathrm{sym}} : \lambda_{\min}(A) \geq 0\big\}, \\ \mathbb{R}^{q\times q}_{\mathrm{sym},>0} &:= \big\{A \in \mathbb{R}^{q\times q}_{\mathrm{sym}} : A \text{ positive definite}\big\} \\ &= \big\{A \in \mathbb{R}^{q\times q}_{\mathrm{sym}} : \lambda_{\min}(A) > 0\big\}. \end{aligned}$$

With $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ we denote the smallest and largest real eigenvalue of a square matrix A. If $A \in \mathbb{R}^{q\times q}$ has only real eigenvalues (e.g. if $A = A^\top$), then $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_q(A)$ are its ordered eigenvalues. The identity matrix in $\mathbb{R}^{q\times q}$ is denoted by I_q.

In the sequel we will introduce further notation and various conditions. For the reader's convenience, these are listed once more at the very end of this paper.

2. Affine and linear equivariance

Affine and linear equivariance are key concepts in connection with estimation of location and scatter. In what follows, let $\mathcal{P}$ be a family of probability distributions on $\mathbb{R}^q$. For $P \in \mathcal{P}$, a vector $a \in \mathbb{R}^q$ and a matrix $B \in \mathbb{R}^{q\times q}_{\rm ns}$ let

$$P^B \;:=\; \mathcal{L}(BX) \quad \text{and} \quad P^{a,B} \;:=\; \mathcal{L}(a+BX) \quad \text{where } X \sim P.$$

Definition 2.1 (Linear equivariance)**.** Suppose that $\mathcal{P}$ is *linear invariant* in the sense that $P^B \in \mathcal{P}$ for arbitrary $P \in \mathcal{P}$ and $B \in \mathbb{R}^{q\times q}_{\rm ns}$. A scatter functional $\mathbf{\Sigma} : \mathcal{P} \to \mathbb{R}^{q\times q}_{{\rm sym},\geq 0}$ is called *linear equivariant* if

$$\mathbf{\Sigma}(P^B) \;=\; B\mathbf{\Sigma}(P)B^\top$$

for arbitrary $P \in \mathcal{P}$ and $B \in \mathbb{R}^{q\times q}_{\rm ns}$.

Definition 2.2 (Affine equivariance)**.** Suppose that $\mathcal{P}$ is *affine invariant* in the sense that $P^{a,B} \in \mathcal{P}$ for arbitrary $P \in \mathcal{P}$, $a \in \mathbb{R}^q$ and $B \in \mathbb{R}^{q\times q}_{\rm ns}$. Consider a location functional $\boldsymbol{\mu} : \mathcal{P} \to \mathbb{R}^q$ and a scatter functional $\mathbf{\Sigma} : \mathcal{P} \to \mathbb{R}^{q\times q}_{{\rm sym},\geq 0}$. These functionals are called *affine equivariant* if

$$\boldsymbol{\mu}(P^{a,B}) \;=\; a + B\boldsymbol{\mu}(P) \quad \text{and} \quad \mathbf{\Sigma}(P^{a,B}) \;=\; B\mathbf{\Sigma}(P)B^\top$$

for arbitrary $P \in \mathcal{P}$, $a \in \mathbb{R}^q$ and $B \in \mathbb{R}^{q\times q}_{\rm ns}$.

These definitions are clearly motivated by the mean vector $\boldsymbol{\mu}(P)$ and covariance matrix $\mathbf{\Sigma}(P)$, where $\mathcal{P}$ consists of all distributions P with finite integral $\int \|x\|^2\, P(dx)$. Whenever we talk about affine or linear equivariant functionals on a set $\mathcal{P}$, we assume tacitly that $\mathcal{P}$ is affine or linear invariant.

Obviously, affine equivariance of a scatter functional $\mathbf{\Sigma}(\cdot)$ implies its linear equivariance. Equivariance properties of location and scatter functionals yield various desirable properties which are summarized in two lemmas below. Let us first recall two symmetry properties of a distribution P:

Definition 2.3 (Spherical and elliptical symmetry)**.** Let X be a random vector with distribution P on $\mathbb{R}^q$.

(i) The distribution P is called spherically symmetric (around 0) if the distributions of X and UX coincide for any orthogonal matrix $U \in \mathbb{R}^{q\times q}$.
(ii) The distribution P is called elliptically symmetric with center $\mu \in \mathbb{R}^q$ and scatter matrix $\Sigma \in \mathbb{R}^{q\times q}_{{\rm sym},>0}$, if the distribution of $\Sigma^{-1/2}(X-\mu)$ is spherically symmetric.

If the distribution P admits a density f, elliptical symmetry with center μ and scatter matrix Σ means that $f(x)$ is a function of the squared Mahalanobis distance $(x-\mu)^\top\Sigma^{-1}(x-\mu)$ only. In particular, if P is spherically symmetric, $f(x)$ depends only on the norm $\|x\|$.

Note that the scatter matrix Σ of an elliptically symmetric distribution is not unique. One could replace Σ with $c\Sigma$ for any $c > 0$.

Lemma 2.4 (Some consequences of linear equivariance)**.** *Let $\mathbf{\Sigma} : \mathcal{P} \to \mathbb{R}^{q\times q}_{{\rm sym},\geq 0}$ be a linear equivariant functional of scatter, and let X be a random vector with distribution $P \in \mathcal{P}$.*

(i) *Let J be a subset of $\{1,2,\ldots,q\}$ with two or more elements. Suppose that the distributions of X and $(X_{\pi(i)})_{i=1}^q$ coincide for any permutation π of $\{1,2,\ldots,q\}$ such that $\pi(i) = i$ whenever $i \notin J$. Then there exist numbers $a = a(P)$ and $b = b(P)$ such that for arbitrary indices $j, k \in J$,*

$$\mathbf{\Sigma}(P)_{jk} \;=\; \begin{cases} a & \text{if } j = k, \\ b & \text{if } j \neq k. \end{cases}$$

(ii) *Suppose that for a given sign vector $s \in \{-1,1\}^q$, the distributions of X and $(s_i X_i)_{i=1}^q$ coincide. Then*

$$\mathbf{\Sigma}(P)_{ij} \;=\; 0 \quad \text{whenever} \quad s_i \neq s_j.$$

(iii) *If P is elliptically symmetric with center $0 \in \mathbb{R}^q$ and scatter matrix $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$, then*

$$\mathbf{\Sigma}(P) \;=\; c(P)\Sigma$$

for some number $c(P) \geq 0$.

Lemma 2.5 (Some consequences of affine equivariance)**.** *Let $\boldsymbol{\mu} : \mathcal{P} \to \mathbb{R}^q$ and $\mathbf{\Sigma} : \mathcal{P} \to \mathbb{R}^{q\times q}_{\mathrm{sym},\geq 0}$ be affine equivariant functionals of location and scatter, respectively, and let X be a random vector with distribution $P \in \mathcal{P}$.*

(i) *Suppose that for a given vector $s \in \{-1,1\}^q$, the distributions of X and $(s_i X_i)_{i=1}^q$ coincide. Then*

$$\boldsymbol{\mu}(P)_i \;=\; 0 \quad \text{whenever} \quad s_i = -1.$$

(ii) *If P is elliptically symmetric with center $\mu \in \mathbb{R}^q$ and scatter matrix $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$, then*

$$\boldsymbol{\mu}(P) \;=\; \mu \quad \text{and} \quad \mathbf{\Sigma}(P) \;=\; c(P)\Sigma$$

for some number $c(P) \geq 0$.

Remark 2.6 (Symmetrization and the block independence property)**.** Suppose that $X \sim P$ may be written as $X = [X_1^\top, X_2^\top]^\top$ with two independent subvectors $X_i \in \mathbb{R}^{q(i)}$, $q(1) + q(2) = q$. Let X' be an independent copy of X. If $\mathbf{\Sigma} : \mathcal{P} \to \mathbb{R}^{q\times q}_{\mathrm{sym},\geq 0}$ is a linear equivariant scatter functional, and if $\tilde{P} := \mathcal{L}(X - X')$ belongs to $\mathcal{P}$,

$$\mathbf{\Sigma}(\tilde{P}) \;=\; \begin{bmatrix} \mathbf{\Sigma}_1(\tilde{P}) & 0 \\ 0 & \mathbf{\Sigma}_2(\tilde{P}) \end{bmatrix}$$

with $\mathbf{\Sigma}_i(\tilde{P}) \in \mathbb{R}^{q(i)\times q(i)}$. This follows from Lemma 2.4 (ii), applied to $\tilde{X} \sim \tilde{P}$ in place of $X \sim P$ and $s_i := 1_{[i\leq q(1)]} - 1_{[i>q(1)]}$. If $\mathcal{P}$ is even affine invariant and $\boldsymbol{\mu} : \mathcal{P} \to \mathbb{R}^q$ an affine equivariant location functional, then $\boldsymbol{\mu}(\tilde{P}) = 0$ by Lemma 2.5.

3. From maximum-likelihood estimation to M-functionals

In this section we describe various estimation problems and the M-functionals which they lead to.

3.1. Estimation in location-scatter families

Let $X_1, X_2, \ldots, X_n$ be independent random vectors with unknown distribution P. As a model for P we consider a location-scatter family constructed as follows: Let $\tilde{f} : [0,\infty) \to [0,\infty)$ satisfy

$$\tilde{c} := \int_{\mathbb{R}^q} \tilde{f}(\|x\|^2)\, dx \;\in\; (0,\infty).$$

For any location parameter $\mu \in \mathbb{R}^q$ and scatter parameter $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$,

$$f_{\mu,\Sigma}(x) \;:=\; \tilde{c}^{-1} \det(\Sigma)^{-1/2} \tilde{f}\big((x-\mu)^\top \Sigma^{-1}(x-\mu)\big)$$

defines a probability density $f_{\mu,\Sigma}$ on $\mathbb{R}^q$. Assuming that P has a density belonging to this family $(f_{\mu,\Sigma})_{\mu,\Sigma}$, a maximum-likelihood estimator of (μ,Σ) is a maximizer $(\widehat{\mu},\widehat{\Sigma})$ of the likelihood function

$$(\mu,\Sigma) \;\mapsto\; \prod_{i=1}^n f_{\mu,\Sigma}(X_i).$$

In other words, $(\widehat{\mu},\widehat{\Sigma})$ minimizes

$$\widehat{L}(\mu,\Sigma) \;:=\; \frac{1}{n}\sum_{i=1}^n \rho\big((X_i-\mu)^\top\Sigma^{-1}(X_i-\mu)\big) + \log\det(\Sigma)$$

with

$$\rho(s) \;:=\; -2\log\tilde{f}(s).$$

The expected value of $\widehat{L}(\mu,\Sigma)$ equals

$$L(\mu,\Sigma,P) \;:=\; \int \rho\big((x-\mu)^\top\Sigma^{-1}(x-\mu)\big)\, P(dx) + \log\det(\Sigma), \tag{3.1}$$

provided this integral exists, and

$$\widehat{L}(\mu,\Sigma) \;=\; L(\mu,\Sigma,\widehat{P})$$

with $\widehat{P}$ denoting the empirical distribution $n^{-1}\sum_{i=1}^n \delta_{X_i}$ of the observations X_i. Consequently we focus on $L(\mu,\Sigma,P)$ for arbitrary distributions P, keeping in mind that P could be a "true" or an empirical distribution.

Suppose that P has a density f which may but need not belong to the model $(f_{\mu,\Sigma})_{\mu,\Sigma}$ and such that $\int f(x)\log f(x)\,dx$ exists in $\mathbb{R}$. Then

$$\begin{aligned} L(\mu,\Sigma,P) - 2\log\tilde{c} \;&=\; -2\int f(x)\log f_{\mu,\Sigma}(x)\,dx \\ &=\; -2\int f(x)\log f(x)\,dx + 2D(f,f_{\mu,\Sigma}) \end{aligned}$$

with the Kullback-Leibler divergence

$$D(f,f_{\mu,\Sigma}) \;:=\; \int f(x)\log\big(f(x)/f_{\mu,\Sigma}(x)\big)\,dx.$$

It is well-known that $D(f, f_{\mu,\Sigma}) \geq 0$ with equality if, and only if, $f = f_{\mu,\Sigma}$ almost everywhere. Thus minimizing $L(\mu, \Sigma, P)$ w.r.t. (μ, Σ) may be viewed as approximating P by one of the densities $f_{\mu,\Sigma}$ in terms of the Kullback-Leibler divergence.

Example 3.1 (Gaussian distributions)**.** Multivariate (nondegenerate) Gaussian distributions correspond to $\tilde{f}(s) := \exp(-s/2)$ and $\tilde{c} := (2\pi)^{q/2}$, i.e. $\rho(s) := s$. Suppose that P has mean vector $\boldsymbol{\mu}(P)$, finite integral $\int \|x\|^2 \, P(dx)$ and nonsingular covariance matrix $\boldsymbol{\Sigma}(P)$. Then

$$\begin{aligned} L(\mu, \Sigma, P) &= \int (x-\mu)^\top \Sigma^{-1} (x-\mu) \, P(dx) + \log\det(\Sigma) \\ &= \int (x - \boldsymbol{\mu}(P))^\top \Sigma^{-1} (x - \boldsymbol{\mu}(P)) \, P(dx) + \log\det(\Sigma) \\ &\quad + \; (\mu - \boldsymbol{\mu}(P))^\top \Sigma^{-1} (\mu - \boldsymbol{\mu}(P)). \end{aligned}$$

Hence for any fixed Σ, the unique minimizer of $\mu \mapsto L(\mu, \Sigma, P)$ equals $\mu = \boldsymbol{\mu}(P)$. Moreover,

$$\begin{aligned} L(\boldsymbol{\mu}(P), \Sigma, P) &= \operatorname{tr}(\Sigma^{-1} \boldsymbol{\Sigma}(P)) + \log\det(\Sigma) \\ &= \operatorname{tr}(\Sigma^{-1} \boldsymbol{\Sigma}(P)) - \log\det(\Sigma^{-1} \boldsymbol{\Sigma}(P)) + \log\det(\boldsymbol{\Sigma}(P)). \end{aligned}$$

Note that $\operatorname{tr}(\Sigma^{-1}\boldsymbol{\Sigma}(P)) - \log\det(\Sigma^{-1}\boldsymbol{\Sigma}(P))$ equals $\operatorname{tr}(B) - \log\det(B)$ with the symmetric matrix $B := \Sigma^{-1/2}\boldsymbol{\Sigma}(P)\Sigma^{-1/2}$. If $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q > 0$ denote the eigenvalues of B, then

$$\operatorname{tr}(B) - \log\det(B) \;=\; \sum_{i=1}^{q} (\lambda_i - \log\lambda_i) \;\geq\; q$$

with equality if, and only if, all eigenvalues λ_i are equal to one, i.e. if $\Sigma = \boldsymbol{\Sigma}(P)$. Thus $(\boldsymbol{\mu}(P), \boldsymbol{\Sigma}(P))$ is the unique minimizer of $L(\cdot, \cdot, P)$.

The range of distributions P for which $L(\mu, \Sigma, P)$ is well-defined in $\mathbb{R}$ for arbitrary (μ, Σ) may become larger if we replace the term $\rho\big((x-\mu)^\top\Sigma^{-1}(x-\mu)\big)$ with a difference

$$\rho\big((x-\mu)^\top\Sigma^{-1}(x-\mu)\big) - \rho\big((x-\mu_o)^\top\Sigma_o^{-1}(x-\mu_o)\big)$$

for some (μ_o, Σ_o). The choice of the latter pair is irrelevant, so we use $\mu_o = 0$ and $\Sigma_o = I_q$, where I_q denotes the unit matrix in $\mathbb{R}^{q\times q}$.

Definition 3.2 (M-functionals of location and scatter)**.** Let $\rho : [0, \infty) \to \mathbb{R}$ be some continuous function. Further let $\mathcal{P}$ be the set of all probability distributions P on $\mathbb{R}^q$ such that

$$L(\mu, \Sigma, P) \;:=\; \int \big[\rho\big((x-\mu)^\top\Sigma^{-1}(x-\mu)\big) - \rho(x^\top x)\big] \, P(dx) + \log\det(\Sigma) \quad (3.2)$$

is well-defined in $\mathbb{R}$ for arbitrary $(\mu, \Sigma) \in \mathbb{R}^q \times \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$.

With $\mathcal{P}_\rho$ we denote the set of all distributions $P \in \mathcal{P}$ such that $L(\cdot, \cdot, P)$ has a unique minimizer $(\boldsymbol{\mu}(P), \boldsymbol{\Sigma}(P))$. This defines an M-functional $\boldsymbol{\mu} : \mathcal{P}_\rho \to \mathbb{R}^q$ of location and an M-functional $\boldsymbol{\Sigma} : \mathcal{P}_\rho \to \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ of scatter.

Affine equivariance The set $\mathcal{P}$ in Definition 3.2 is affine invariant. Indeed, if $X \sim P \in \mathcal{P}$ and $X' := a + BX \sim P^{a,B}$, then elementary calculations show that

$$\begin{aligned}
&\rho\big((X'-\mu')^\top \Sigma'^{-1}(X'-\mu')\big) - \rho(X'^\top X') \\
&\quad = \Big[\rho\big((X-\mu)^\top \Sigma^{-1}(X-\mu)\big) - \rho(X^\top X)\Big] \\
&\qquad - \Big[\rho\big((X-\mu'')^\top \Sigma''^{-1}(X-\mu'')\big) - \rho(X^\top X)\Big],
\end{aligned}$$

where $\mu' := a + B\mu$, $\Sigma' := B\Sigma B^\top$ and $\mu'' := -B^{-1}a$, $\Sigma'' := (B^\top B)^{-1}$. Since $\log\det(\Sigma') = \log\det(\Sigma) + 2\log|\det(B)| = \log\det(\Sigma) - \log\det(\Sigma'')$, we arrive at the key equation

$$L\big(a + B\mu, B\Sigma B^\top, P^{a,B}\big) \;=\; L(\mu, \Sigma, P) + c(a, B, P) \tag{3.3}$$

with $c(a, B, P) := -L\big(-B^{-1}a, (B^\top B)^{-1}, P\big)$. In particular, the set $\mathcal{P}_\rho$ is affine invariant, and the M-functionals $\boldsymbol{\mu}(\cdot)$, $\boldsymbol{\Sigma}(\cdot)$ are affine equivariant.

Example 3.3 (Multivariate t-distributions)**.** The multivariate student-distributions are generated by $\tilde{f}(s) := (\nu+s)^{-(\nu+q)/2}$ for a fixed parameter $\nu > 0$, the "degrees of freedom", and $\tilde{c} = \nu^{-\nu/2}\pi^{q/2}\Gamma(\nu/2)/\Gamma((\nu+q)/2)$. Here

$$\rho(s) \;=\; (\nu+q)\log(\nu+s).$$

With this choice of ρ, definition (3.2) yields

$$\begin{aligned}
&L_\nu(\mu, \Sigma, P) \\
&\quad = (\nu+q)\int \log\Big(\frac{\nu + (x-\mu)^\top\Sigma^{-1}(x-\mu)}{\nu + \|x\|^2}\Big)\, P(dx) + \log\det(\Sigma).
\end{aligned} \tag{3.4}$$

Since the integrand is continuous and bounded on $\mathbb{R}^q$ for any fixed (μ, Σ), the set $\mathcal{P}$ is just the set of *all* probability distributions on $\mathbb{R}^q$. In later sections we shall derive a precise description of the corresponding subset $\mathcal{P}_\rho$.

3.2. Tyler's (1987) M-functional of scatter and more

A maximum-likelihood estimator for directional data Tyler (1987a; 1987b) introduced a particular M-estimator of scatter which may be motivated as follows: Suppose that $X_1, X_2, \ldots, X_n$ are independent random vectors with possibly different distributions $P_1, P_2, \ldots, P_n$ on $\mathbb{R}^q$. However, suppose that each P_i satisfies $P_i(\{0\}) = 0$ and is elliptically symmetric with center 0 and a common scatter matrix Σ. This assumption means that $X_i = R_i B U_i$ with $B := \Sigma^{1/2}$ and $2n$ stochastically independent random variables $R_1, R_2, \ldots, R_n > 0$ and $U_1, U_2, \ldots, U_n$ uniformly distributed on the unit sphere $\mathbb{S}^{q-1}$ of $\mathbb{R}^q$. In particular, the directional vectors $V_i := \|X_i\|^{-1}X_i = \|BU_i\|^{-1}BU_i$ are independent and identically distributed random vectors. One can show that V_i possesses a so

called angular central Gaussian distribution, i.e. its distribution is absolutely continuous with respect to the uniform distribution on $\mathbb{S}^{q-1}$ with density

$$g_\Sigma(v) \;:=\; \det(\Sigma)^{-1/2}(v^\top \Sigma^{-1} v)^{-q/2},$$

see e.g. Watson (1983). Consequently, a maximum-likelihood estimator for Σ is given by a maximizer of the target function $L(\Sigma, \widehat{P})$ over all matrices $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$, where $\widehat{P}$ is again the empirical distribution of the X_i, and

$$L(\Sigma, P) \;:=\; q \int \log\Bigl(\frac{x^\top \Sigma^{-1} x}{x^\top x}\Bigr)\, P(dx) + \log\det(\Sigma) \tag{3.5}$$

for any distribution P on $\mathbb{R}^q$ with $P(\{0\}) = 0$. Note that $L(\Sigma, P) = L(c\Sigma, P)$ for any $c > 0$. To achieve uniqueness of a minimizer, we have to impose an additional constraint, e.g.

$$\det(\Sigma) \;\stackrel{!}{=}\; 1,$$

following Paindaveine's (2008) advice.

Estimation of proportional covariance matrices Suppose that one observes independent random matrices $S_1, S_2, \ldots, S_K \in \mathbb{R}^{q\times q}_{\mathrm{sym},\geq 0}$, where S_i has a Wishart distribution $\mathcal{W}_q(c_i\Sigma, m_i)$. The degrees of freedom, $m_1, m_2, \ldots, m_K$, are given, while $c_1, c_2, \ldots, c_K > 0$ and $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ are unknown parameters.

As an explicit example, suppose that we observe independent random vectors $X_{ij} \in \mathbb{R}^q$ for $1 \leq i \leq K$ and $1 \leq j \leq n_i$, where $n_i = m_i + 1 \geq 2$ and

$$X_{ij} \;\sim\; \mathcal{N}_q(\mu_i, c_i\Sigma)$$

with unknown means $\mu_i \in \mathbb{R}^q$. With $\bar{X}_i := n_i^{-1}\sum_{j=1}^{n_i} X_{ij}$, the standard estimator of μ_i, it is well-known that

$$S_i \;:=\; \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)^\top \;\sim\; \mathcal{W}_q(c_i\Sigma, m_i).$$

Recalling that $\mathcal{W}_q(\Gamma, m)$ stands for the distribution of $\sum_{j=1}^m Y_i Y_i^\top$ with independent random vectors $Y_1, \ldots, Y_m \sim \mathcal{N}_q(0, \Gamma)$, the log-likelihood function times -2 may be written as

$$\sum_{i=1}^K \bigl(c_i^{-1}\,\mathrm{tr}(\Sigma^{-1}S_i) + qm_i \log c_i + m_i \log\det(\Sigma)\bigr). \tag{3.6}$$

Minimization of this function was treated by Flury (1986), Eriksen (1987) and Jensen and Johansen (1987). The proposed algorithms rely on the fact that (3.6), as a function of the two arguments Σ and $c = (c_i)_{i=1}^K$, is easily minimized if one of the two arguments is fixed. For fixed Σ, the unique minimizer is

$$c(\Sigma) \;:=\; q^{-1}\bigl(m_i^{-1}\,\mathrm{tr}(\Sigma^{-1}S_i)\bigr)_{i=1}^K,$$

whereas for fixed c, the unique minimizer is

$$\Sigma(c) \;:=\; m_+^{-1} \sum_{i=1}^{K} c_i^{-1} S_i$$

with $m_+ := \sum_{i=1}^{K} m_i$. If focusing on the estimation of the matrix parameter Σ, we may plug $c(\Sigma)$ into (3.6) and try to minimize the resulting function of Σ. Up to an additive term and a scaling factor m_+^{-1}, the latter function equals

$$q \sum_{i=1}^{K} \frac{m_i}{m_+} \log\Big(\frac{\operatorname{tr}(\Sigma^{-1} S_i)}{\operatorname{tr}(S_i)}\Big) + \log\det(\Sigma). \tag{3.7}$$

Again one should impose some constraint such as $\det(\Sigma) \stackrel{!}{=} 1$ to avoid non-uniqueness of the minimizer.

A generalized setting Note the similarity between (3.5) and (3.7). Consider the distribution Q of the random matrix $XX^\top$, where $X \sim P$. Then $L(\Sigma, P)$ in (3.5) may be rewritten as

$$q \int \log\Big(\frac{\operatorname{tr}(\Sigma^{-1} M)}{\operatorname{tr}(M)}\Big)\, Q(dM) + \log\det(\Sigma),$$

where M corresponds to $xx^\top$ with $x \in \mathbb{R}^q$. But (3.7) is also of this form, this time with the random distribution

$$\widehat{Q} \;:=\; \sum_{i=1}^{K} \frac{m_i}{m_+}\, \delta_{S_i}$$

in place of Q. These considerations motivate the following definition.

Definition 3.4 (Generalized version of Tyler's M-functional of scatter)**.** For a distribution Q on $\mathbb{R}^{q\times q}_{\mathrm{sym},\geq 0} \setminus \{0\}$ and $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ we define

$$L_0(\Sigma, Q) \;:=\; q \int \log\Big(\frac{\operatorname{tr}(\Sigma^{-1} M)}{\operatorname{tr}(M)}\Big)\, Q(dM) + \log\det(\Sigma).$$

If $L_0(\cdot, Q)$ has a unique minimizer Σ satisfying $\det(\Sigma) = 1$, then we denote it with $\Sigma_0(Q)$.

3.3. Symmetrizations of arbitrary order

For $k \geq 2$ vectors $x_1, \ldots, x_k \in \mathbb{R}^q$ we define their sample covariance matrix as

$$S(x_1, \ldots, x_k) \;:=\; \frac{1}{k-1} \sum_{i=1}^{k} (x_i - \bar{x})(x_i - \bar{x})^\top$$

with $\bar{x} := k^{-1} \sum_{i=1}^{k} x_i$. If $X_1, X_2, \ldots, X_n$ are independent random vectors with distribution P such that $\int \|x\|^2 P(dx) < \infty$, then $S(X_1, X_2, \ldots, X_n)$ is an unbiased estimator of the covariance matrix of P. Elementary calculations show

that

$$\begin{aligned} S(X_1, X_2, \ldots, X_n) &= \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} 2^{-1}(X_i - X_j)(X_i - X_j)^\top \\ &= \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} S(X_i, X_j). \end{aligned}$$

More generally, for $2 \le k \le n$,

$$S(X_1, X_2, \ldots, X_n) = \binom{n}{k}^{-1} \sum_{1 \le i_1 < \cdots < i_k \le n} S(X_{i_1}, \ldots, X_{i_k}).$$

Instead of taking the average of all sample covariance matrices $S(X_{i_1}, \ldots, X_{i_k})$ one could apply Tyler's generalized M-functional of scatter (Definition 3.4) or other functionals of scatter to the random distribution

$$\binom{n}{k}^{-1} \sum_{1 \le i_1 < \cdots < i_k \le n} \delta_{S(X_{i_1}, \ldots, X_{i_k})}$$

on $\mathbb{R}^{q \times q}_{\mathrm{sym},\ge 0}$, a measure-valued U-statistic (cf. Hoeffding, 1948). For $k = 2$ this approach was proposed by Dümbgen (1998). Apart from the higher computational complexity, trying $k \ge 3$ is tempting.

3.4. Simultaneous symmetrization in several samples

Suppose we observe independent random vectors $X_{ij} \in \mathbb{R}^q$, where $i = 1, 2, \ldots, K$ and $j = 1, 2, \ldots, n_i$, $n_i \ge 2$. Suppose that X_{ij} has an unknown elliptically symmetric distribution P_i with center $\mu_i \in \mathbb{R}^q$ and a common scatter matrix $\Sigma \in \mathbb{R}^{q \times q}_{\mathrm{sym},>0}$. In case of $P_i = \mathcal{N}_q(\mu_i, \Sigma)$ one could estimate Σ by the usual pooled covariance matrix

$$\begin{aligned} \widehat{\Sigma} &= \frac{1}{n_+ - K} \sum_{i=1}^{K} (n_i - 1) S(X_{i1}, X_{i2}, \ldots, X_{in_i}) \\ &= \frac{2}{n_+ - K} \sum_{i=1}^{K} \frac{1}{n_i} \sum_{1 \le j < \ell \le n_i} S(X_{ij}, X_{i\ell}). \end{aligned}$$

Alternatively, one could estimate Σ by a minimizer of (3.7). But in case of potentially heavy-tailed distributions P_i, it might be even better to apply Tyler's generalized M-functional of scatter (Definition 3.4) or other functionals of scatter to the random distribution

$$\frac{2}{n_+ - K} \sum_{i=1}^{K} \frac{1}{n_i} \sum_{1 \le j < \ell \le n_i} \delta_{S(X_{ij}, X_{i\ell})}$$

on $\mathbb{R}^{q \times q}_{\mathrm{sym},\ge 0}$.

The resulting scatter estimator $\widehat{\Sigma}$ could be used, for instance, in the context of nearest-neighbor classification to define a data-driven Mahalanobis distance $\widehat{d}(x,y) := \|\widehat{\Sigma}^{-1/2}(x-y)\|$ between vectors $x, y \in \mathbb{R}^q$.

4. *M*-functionals of scatter

In this section we consider M-functionals of scatter only. That means, when thinking about a distribution on $\mathbb{R}^q$, we assume that it has a given center $\mu = 0$. In view of the considerations in the preceding section, however, we consider distributions Q on $\mathbb{R}^{q\times q}_{\mathrm{sym},\geq 0}$. Two particular examples for Q are

$$Q^1(P) \;:=\; \mathcal{L}(XX^\top) \tag{4.1}$$

and

$$Q^k(P) \;:=\; \mathcal{L}\big(S(X_1, X_2, \ldots, X_k)\big), \quad k \geq 2, \tag{4.2}$$

for independent, identically distributed random vectors $X, X_1, X_2, \ldots, X_k$ with distribution P on $\mathbb{R}^q$.

4.1. Definitions and basic properties

Definition 4.1 (A log-likelihood type criterion)**.** For a given "loss function" $\rho : [0,\infty) \to \mathbb{R}$ we define

$$L_\rho(\Sigma, Q) \;:=\; \int \big[\rho(\mathrm{tr}(\Sigma^{-1}M)) - \rho(\mathrm{tr}(M))\big]\, Q(dM) + \log\det\Sigma$$

for $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$, provided that the integral exists in $\mathbb{R}$.

Assumptions on ρ and Q Throughout we assume that ρ is continuously differentiable on $(0,\infty)$ with derivative $\rho' > 0$. Moreover, we assume that

$$\psi(s) \;:=\; s\rho'(s).$$

is non-decreasing in $s > 0$.

Case 0 For $s > 0$ let

$$\rho(s) \;:=\; q\log(s),$$

so $\rho'(s) = q/s$ and $\psi(s) = q$. Here we assume that $Q(\{0\}) = 0$.

Case 1 We assume that ψ is strictly increasing on $(0,\infty)$ with limits $\psi(0) = 0$ and $\psi(\infty) \in (q,\infty]$. Here we assume that

$$\int \psi(\lambda\,\mathrm{tr}(M))\, Q(dM) \;<\; \infty \quad \text{for any } \lambda \geq 1, \tag{4.3}$$

which is obviously true in case of $\psi(\infty) < \infty$.

Remark 4.2. Note that Tyler's generalized M-functional of scatter (Definition 3.4) corresponds to Case 0 above. In Case 1, if $Q = Q^1(P)$ as in (4.1), then $L_\rho(\cdot, Q)$ corresponds to the log-likelihood function $L(0, \Sigma, P)$ for an elliptical model with $\tilde{f}(s) := \exp(-\rho(s)/2)$. Note that for $0 < s_o < s$,

$$\rho(s) \;=\; \rho(s_o) + \int_{s_o}^{s} \psi(t) t^{-1}\, dt \;\begin{cases} \leq\; \rho(s_o) + \psi(\infty) \log(s/s_o), \\ \geq\; \rho(s_o) + \psi(s_o) \log(s/s_o). \end{cases}$$

This implies that

$$\int_{\mathbb{R}^q} \exp\bigl(-\rho(\|x\|^2)/2\bigr)\, dx \;=\; C_q \int_0^\infty \exp\bigl(-\rho(s)/2 + (q/2-1)\log(s)\bigr)\, ds$$

is finite if, and only if, $\psi(\infty) > q$.

Remark 4.3. Several authors require in addition ρ' to be non-increasing on $(0, \infty)$. Then

$$\psi(\lambda s) \;=\; \lambda s \rho'(\lambda s) \;\leq\; \lambda \psi(s)$$

for any $s > 0$ and $\lambda \geq 1$, whence (4.3) is equivalent to

$$\int \psi(\mathrm{tr}(M))\, Q(dM) \;<\; \infty.$$

Example 4.4 (Multivariate t-distributions)**.** For $\nu \geq 0$ let

$$\rho(s) = \rho_{\nu,q}(s) \;:=\; (\nu + q) \log(\nu + s).$$

In case of $\nu > 0$, $L_\rho(\Sigma, Q)$ in Definition 4.1 may be viewed as a generalization of $L_\nu(0, \Sigma, P)$ in (3.4). Here $\rho'(s) = (\nu + q)/(\nu + s)$ is strictly decreasing and $\psi(s) = (\nu + q)s/(\nu + s)$ is strictly increasing in $s \geq 0$. Moreover, $\psi(0) = 0$ and $\psi(\infty) = \nu + q$.

Example 4.5 (Multivariate elliptical Weibull-distributions)**.** For a fixed $\gamma > 0$ and $s \geq 0$ let $\rho(s) := s^\gamma$. Then $\rho'(s) = \gamma s^{\gamma-1}$ and $\psi(s) := \gamma s^\gamma$. Here $L_\rho(\Sigma, Q)$ corresponds to elliptically symmetric distributions with center 0 that are generated by $\tilde{f}(s) := \exp(-s^\gamma/2)$. In this situation (4.3) means that

$$\int \mathrm{tr}(M)^\gamma\, Q(dM) \;<\; \infty,$$

and in setting (4.1) this is equivalent to

$$\int \|x\|^{2\gamma}\, P(dx) \;<\; \infty.$$

Example 4.6. Another example, suggested to us by David Tyler, is given by

$$\rho(s) \;:=\; (\nu + q) \log(1 + s^2)/2$$

for $s \geq 0$ with some parameter $\nu > 0$. Here $\rho'(s) = (\nu + q)s/(1 + s^2)$, and $\psi(s) = (\nu + q)s^2/(1 + s^2)$ is strictly increasing in $s \geq 0$ with $\psi(0) = 0$ and $\psi(\infty) = \nu + q$.

Existence of L_ρ The functional $L_\rho(\cdot, P) : \mathbb{R}^{q\times q}_{\mathrm{sym},>0} \to \mathbb{R}$ is well-defined in Cases 0 and 1. This will be derived from the following two elementary inequalities which will be used several times:

Lemma 4.7. *For $M \in \mathbb{R}^{q\times q}_{\mathrm{sym},\geq 0}$ and $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$,*

$$\lambda_{\min}(A)\operatorname{tr}(M) \;\leq\; \operatorname{tr}(AM) \;\leq\; \lambda_{\max}(A)\operatorname{tr}(M).$$

Lemma 4.8. *For arbitrary $s, t > 0$,*

$$\psi(s)\log(t/s) \;\leq\; \rho(t) - \rho(s) \;\leq\; \psi(t)\log(t/s).$$

If ρ' is non-increasing on $(0,\infty)$, then

$$\rho'(t)(t-s) \;\leq\; \rho(t) - \rho(s) \;\leq\; \rho'(s)(t-s).$$

It follows from Lemma 4.7 that for arbitrary $M \in \mathbb{R}^{q\times q}_{\mathrm{sym},\geq 0}$ and $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$,

$$\lambda_{\max}(\Sigma)^{-1}\operatorname{tr}(M) \;\leq\; \operatorname{tr}(\Sigma^{-1}M) \;\leq\; \lambda_{\min}(\Sigma)^{-1}\operatorname{tr}(M).$$

Combining these inequalities in case of $M \neq 0$ with Lemma 4.8, applied to $\{s,t\} = \{\operatorname{tr}(M), \operatorname{tr}(\Sigma^{-1}M)\}$, yields the inequality

$$\big|\rho(\operatorname{tr}(\Sigma^{-1}M)) - \rho(\operatorname{tr}(M))\big| \;\leq\; \psi\big(\lambda_*(\Sigma)\operatorname{tr}(M)\big)\log(\lambda_*(\Sigma))$$

with $\lambda_*(\Sigma) = \max\{\lambda_{\min}(\Sigma)^{-1}, \lambda_{\max}(\Sigma)\}$, and the right hand side is integrable with respect to Q by assumption (4.3).

Linear equivariance For a nonsingular matrix $B \in \mathbb{R}^{q\times q}$ let

$$Q^B \;:=\; \mathcal{L}(BSB^\top) \quad\text{and}\quad Q_B \;:=\; \mathcal{L}(B^{-1}SB^{-\top}) \quad\text{with}\quad S \sim Q,$$

where $B^{-\top} := (B^{-1})^\top = (B^\top)^{-1}$. Then one can easily verify that for arbitrary $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$,

$$\begin{aligned} L_\rho(B\Sigma B^\top, Q^B) - L_\rho(BB^\top, Q^B) &= L_\rho(\Sigma, Q), \\ L_\rho(B\Sigma B^\top, Q) - L_\rho(BB^\top, Q) &= L_\rho(\Sigma, Q_B). \end{aligned} \tag{4.4}$$

Let $\mathcal{Q}_\rho$ denote the set of all distributions Q as described in Cases 0 and 1 such that $L_\rho(\cdot, Q)$ has a unique minimizer in

$$\begin{cases} \{\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0} : \det(\Sigma) = 1\} & \text{in Case 0,} \\ \mathbb{R}^{q\times q}_{\mathrm{sym},>0} & \text{in Case 1.} \end{cases}$$

This minimizer is denoted by $\boldsymbol{\Sigma}_\rho(Q)$. Then $\mathcal{Q}_\rho$ is linear invariant and $\boldsymbol{\Sigma}_\rho$ is linear equivariant in the sense that $Q^B \in \mathcal{Q}_\rho$ and

$$\boldsymbol{\Sigma}(Q^B) \;=\; \begin{cases} \det(BB^\top)^{-1/q} B\boldsymbol{\Sigma}_\rho(Q)B^\top & \text{in Case 0} \\ B\boldsymbol{\Sigma}_\rho(Q)B^\top & \text{in Case 1} \end{cases}$$

for all $Q \in \mathcal{Q}_\rho$ and $B \in \mathbb{R}^{q\times q}_{\mathrm{ns}}$.

4.2. Existence and uniqueness of an optimizer

The question of existence and uniqueness of minimizers of $L_\rho(\cdot, Q)$ is closely related to the mass which Q puts on special linear subspaces of $\mathbb{R}^{q\times q}_{\mathrm{sym}}$. We define

$$\mathcal{V}_q \;:=\; \{\mathbb{V} : \mathbb{V} \text{ is a linear subspace of } \mathbb{R}^q\}.$$

Then for $\mathbb{V} \in \mathcal{V}_q$, we consider

$$\mathbb{M}(\mathbb{V}) \;:=\; \big\{M \in \mathbb{R}^{q\times q}_{\mathrm{sym},\geq 0} : M\mathbb{R}^q \subset \mathbb{V}\big\},$$

a linear subspace of $\mathbb{R}^{q\times q}_{\mathrm{sym}}$ with dimension $\dim(\mathbb{M}(\mathbb{V})) = \dim(\mathbb{V})(\dim(\mathbb{V})+1)/2$. Another object of interest is the matrix

$$\begin{aligned}\Psi_\rho(\Sigma, Q) \;&:=\; \int \rho'(\mathrm{tr}(\Sigma^{-1}M))M\, Q(dM)\\ &=\; \int \psi(\mathrm{tr}(\Sigma^{-1}M))\,\mathrm{tr}(\Sigma^{-1}M)^{-1}M\, Q(dM),\end{aligned}$$

where the integrands are interpreted as $0 \in \mathbb{R}^{q\times q}$ if $M = 0$. It will turn out that the following conditions play the key role for the existence of a unique minimizer $\Sigma_\rho(Q)$.

Condition 0 We assume that

$$Q(\mathbb{M}(\mathbb{V})) \;<\; \frac{\dim(\mathbb{V})}{q} \quad \text{for all } \mathbb{V} \in \mathcal{V}_q \text{ with } 1 \leq \dim(\mathbb{V}) < q. \tag{4.5}$$

Condition 1 We assume that

$$Q(\mathbb{M}(\mathbb{V})) \;<\; \frac{\psi(\infty) - q + \dim(\mathbb{V})}{\psi(\infty)} \quad \text{for all } \mathbb{V} \in \mathcal{V}_q \text{ with } 0 \leq \dim(\mathbb{V}) < q. \tag{4.6}$$

In case of $\psi(\infty) = \infty$ the fraction on the right hand side of (4.6) is interpreted as 1.

Theorem 4.9. *A matrix $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ minimizes $L_\rho(\cdot, Q)$ if, and only if,*

$$\Psi_\rho(\Sigma, Q) \;=\; \Sigma. \tag{4.7}$$

In Case 0, $L_\rho(\cdot, Q)$ possesses a unique minimizer with determinant 1 if, and only if, Condition 0 is satisfied.

In Case 1, $L_\rho(\cdot, P)$ possesses a unique minimizer if, and only if, Condition 1 is satisfied.

Our proof of Theorem 4.9 is based on an in-depth analysis of the mapping $L_\rho(\cdot, Q)$ in Section 5. In particular it will turn out that the fixed-point equation (4.7) is equivalent to $L_\rho(\cdot, Q)$ having gradient 0 at Σ. With Theorem 4.9 at hand we may redefine the family $\mathcal{Q}_\rho$ as follows:

In Case 0, $\mathcal{Q}_\rho$ consists of all probability distributions Q on $\mathbb{R}^{q\times q}_{\mathrm{sym},\geq 0}$ satisfying Condition 0 and $Q(\{0\}) = 0$.

In Case 1, $\mathcal{Q}_\rho$ consists of all probability distributions Q on $\mathbb{R}^{q\times q}_{\mathrm{sym},\geq 0}$ satisfying Condition 1 and $\int \psi(\lambda\,\mathrm{tr}(M))\, Q(dM) < \infty$ for any $\lambda \geq 1$.

Let us comment now on these conditions in two special settings.

The setting (4.1) If $Q = Q^1(P) = \mathcal{L}(XX^\top)$ with a random vector $X \sim P$, then $Q(\{0\}) = P(\{0\})$, and $\int \psi(\lambda \operatorname{tr}(M))\, Q(dM) = \int \psi(\lambda\|x\|^2)\, P(dx)$. Moreover,

$$Q(\mathbb{M}(\mathbb{V})) \;=\; P(\mathbb{V}).$$

Hence Conditions 0 and 1 coincide with the known conditions from the literature on M-estimation of scatter. In particular, a unique minimizer $\mathbf{\Sigma}_\rho(Q)$ is well-defined if P is smooth in the sense that

$$P(\mathbb{V}) \;=\; 0 \quad \text{for any } \mathbb{V} \in \mathcal{V}_q \text{ with } \dim(\mathbb{V}) < q \tag{4.8}$$

and satisfies $\int \psi(\lambda\|x\|^2)\, P(dx) < \infty$ for any $\lambda \geq 1$.

Now consider the empirical distribution

$$\widehat{Q}^1 \;:=\; n^{-1} \sum_{i=1}^{n} \delta_{X_i X_i^\top}$$

with $n \geq q$ independent random vectors $X_1, X_2, \ldots, X_n \sim P$. This is an unbiased estimator of $Q^1(P)$. In Section 8 we will apply Theorem 4.9 to $\widehat{Q}^1$ and prove the following result:

Lemma 4.10. *Suppose that P is smooth in the sense of (4.8). Then $\mathbf{\Sigma}(\widehat{Q}^1)$ is well-defined with probability one, provided that*

$$n \;\geq\; \begin{cases} q+1 & \text{in Case 0,} \\ q & \text{in Case 1.} \end{cases}$$

This result is based on the fact that in case of (4.8), q independent random vectors with distribution P are linearly independent almost surely.

The setting (4.2) Let $Q = Q^k(P) = \mathcal{L}\big(S(X_1, X_2, \ldots, X_k)\big)$ with $k \geq 2$ independent random vectors $X_1, X_2, \ldots, X_k \sim P$. Here $Q(\{0\}) = 0$ if, and only if, P has no atoms, i.e.

$$P(\{x\}) \;=\; 0 \quad \text{for all } x \in \mathbb{R}^q.$$

Note also that $\operatorname{tr}(S(X_1, X_2, \ldots, X_k)) \leq (k-1)^{-1} \sum_{i=1}^k \|X_i\|^2$, so

$$\begin{aligned} \psi\big(\lambda \operatorname{tr}(S(X_1, X_2, \ldots, X_k))\big) \;&\leq\; \psi\Big(\lambda(1-1/k)^{-1} \max_{1\leq i\leq k} \|X_i\|^2\Big) \\ &\leq\; \sum_{i=1}^{k} \psi\big(\lambda(1-1/k)^{-1}\|X_i\|^2\big) \end{aligned}$$

and

$$\int \psi(\lambda \operatorname{tr}(M))\, Q(dM) \;\leq\; k \int \psi\big(\lambda(1-1/k)^{-1}\|x\|^2\big)\, P(dx). \tag{4.9}$$

Moreover, according to Lemma 8.1 in Section 8,

$$S(X_1, X_2, \ldots, X_k)\,\mathbb{R}^q \;=\; \mathrm{span}(X_2 - X_1, \ldots, X_k - X_1).$$

Hence

$$\begin{aligned} Q(\mathbb{M}(\mathbb{V})) &= \mathbb{P}\big(\mathrm{span}(X_2 - X_1, \ldots, X_k - X_1) \subset \mathbb{V}\big) \\ &= \mathbb{P}(X_2 - X_1, \ldots, X_k - X_1 \in \mathbb{V}) \\ &= \int P(x + \mathbb{V})^{k-1}\, P(dx) \\ &= \sum_{w \in \mathbb{V}^\perp} P(w + \mathbb{V})^k. \end{aligned}$$

In particular, $\mathbf{\Sigma}_\rho(Q)$ is well-defined if P is smooth in the sense that

$$P(H) \;=\; 0 \quad \text{for any hyperplane } H \subset \mathbb{R}^q, \tag{4.10}$$

and if $\int \psi(\lambda\|x\|^2)\, P(dx) < \infty$ for arbitrary $\lambda \geq 1$. (A hyperplane is a set of the form $w + \mathbb{V}$ with $w \in \mathbb{R}^q$, $\mathbb{V} \in \mathcal{V}_q$, $\dim(\mathbb{V}) = q - 1$.)

Now consider the empirical distribution

$$\widehat{Q}^k \;:=\; \binom{n}{k}^{-1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} \delta_{S(X_{i_1}, \ldots, X_{i_k})}$$

for some $k \geq 2$ and $n \geq k$ independent random vectors $X_1, X_2, \ldots, X_n \sim P$. Note that $\widehat{Q}^k$ is an unbiased estimator of $Q^k(P)$. In Section 8 we'll prove the following result:

Lemma 4.11. *Suppose that P is smooth in the sense of (4.10). Then $\mathbf{\Sigma}(\widehat{Q}^k)$ is well-defined almost surely, provided that $n \geq q + 1$.*

Estimation of proportional covariance matrices As in Section 3.2 consider

$$\widehat{Q} \;=\; \sum_{i=1}^K \frac{m_i}{m_+}\, \delta_{S_i}$$

with independent random matrices $S_i \sim \mathcal{W}_q(c_i\Sigma, m_i)$. Let $S_i = c_i \sum_{j=1}^{m_i} Y_{ij} Y_{ij}^\top$ with independent random vectors $Y_{ij} \sim \mathcal{N}_q(0, \Sigma)$, $1 \leq i \leq K$, $1 \leq j \leq m_i$. Then one can easily show that

$$S_i\,\mathbb{R}^q \;=\; \mathrm{span}(Y_{ij} : 1 \leq j \leq m_i).$$

Thus with similar arguments as in the proof of Lemma 4.10 one can show that with probability one,

$$\widehat{Q}(\mathbb{M}(\mathbb{V})) \;\leq\; \frac{1}{m_+} \sum_{i=1}^K \sum_{j=1}^{m_i} 1_{[Y_{ij} \in \mathbb{V}]} \;\leq\; \frac{\dim(\mathbb{V})}{m_+}$$

for arbitrary $\mathbb{V} \in \mathcal{V}_q$ with $\dim(\mathbb{V}) < q$. Hence $\mathbf{\Sigma}_\rho(\widehat{Q})$ is well-defined in Case 0 almost surely, provided that

$$m_+ \;\geq\; q + 1.$$

4.3. A fixed-point algorithm

Suppose that ρ satisfies the additional constraint that ρ' is non-increasing on $(0,\infty)$. In this case one can use the fixed-point equation (4.7) to calculate $\mathbf{\Sigma}_\rho(Q)$ numerically. Recall that $\Sigma_* \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ minimizes $L_\rho(\cdot,Q)$ if, and only if, $\Psi_\rho(\Sigma_*,Q)=\Sigma_*$, according to Theorem 4.9. This fixed-point equation implies that

$$\Psi_\rho(\Sigma,Q) \;\in\; \mathbb{R}^{q\times q}_{\mathrm{sym},>0} \quad \text{for arbitrary } \Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}.$$

For otherwise we could find a vector $v \in \mathbb{R}^q \setminus \{0\}$ such that

$$0 \;=\; v^\top \Psi_\rho(\Sigma,Q)v \;=\; \int \rho'(\mathrm{tr}(\Sigma^{-1}M))v^\top M v\, Q(dM).$$

But then $v^\top M v = 0$ for almost all M w.r.t. Q, i.e. $Q(\mathbb{M}(v^\top))=1$. This would yield the contradiction $0 < v^\top \Sigma_* v = v^\top \Psi_\rho(\Sigma_*,Q)v = 0$. It would also contradict Condition 0 and 1.

Iterating the mapping $\Psi_\rho(\cdot,Q)$ yields a sequence converging to a positive multiple of $\mathbf{\Sigma}_\rho(Q)$ in Case 0 and to $\mathbf{\Sigma}_\rho(Q)$ in Case 1:

Lemma 4.12 (Convergence of a fixed-point algorithm)**.** *Suppose that Q fulfills Condition 0 in Case 0 and Condition 1 in Case 1, and let ρ' be non-increasing on $(0,\infty)$. For any starting point $\Sigma_0 \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$, define inductively*

$$\Sigma_k \;:=\; \Psi_\rho(\Sigma_{k-1},Q)$$

for $k = 1,2,3,\ldots$. Then the sequence $(\Sigma_k)_{k\geq 0}$ converges to a solution of the fixed-point equation (4.7).

A key ingredient for proving this lemma is the following inequality. It may be viewed as a special case of a wellkown inequality for the EM algorithm by Dempster et al. (1977). For the precise connection between variations of the present fixed-point algorithm and the EM algorithm we refer to Arslan et al. (1995) and Arslan and Kent (1998).

Lemma 4.13. *Suppose that ρ' is non-increasing on $(0,\infty)$. Let Q be a probability distribution on $\mathbb{R}^{q\times q}_{\mathrm{sym},\geq 0}$ such that $Q(\mathbb{M}(v^\top)) < 1$ for any $v \in \mathbb{R}^q \setminus \{0\}$ and $\int \psi(\mathrm{tr}(M))\, Q(dM) < \infty$. Then for any $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$,*

$$L_\rho(\Psi_\rho(\Sigma,Q),Q) \;<\; L_\rho(\Sigma,Q)$$

unless $\Psi_\rho(\Sigma,Q)=\Sigma$.

5. Analytical properties of the criterion function

The results in the previous section can be derived from an in-depth analysis of the function $L_\rho(\cdot,Q)$. As mentioned in the introduction, we utilize matrix exponentials which are reviewed in the next subsection. Then we derive differentiability, a convexity property and coercivity of $L_\rho(\cdot,Q)$ under certain conditions. In the last subsection we derive second order Taylor expansions of $L_\rho(\cdot,Q)$ which are needed later on.

5.1. The exponential transform of matrices

The exponential transform on $\mathbb{R}^{q\times q}$ For an arbitrary matrix $A \in \mathbb{R}^{q\times q}$, its exponential transform

$$\exp(A) \;:=\; \sum_{k=0}^{\infty} \frac{A^k}{k!}$$

is well-defined in $\mathbb{R}^{q\times q}$, satisfying the inequalities $\|\exp(A)\| \le e^{\|A\|}$ and

$$\Big\|\sum_{k=\ell}^{\infty} \frac{A^k}{k!}\Big\| \;\le\; e^{\|A\|}\|A\|^\ell/\ell! \quad \text{for } \ell \ge 1.$$

If $A, B \in \mathbb{R}^{q\times q}$ are interchangeable in the sense that $AB = BA$, the familiar equation $\exp(A+B) = \exp(A)\exp(B) = \exp(B)\exp(A)$ is valid. In particular, $\exp(A)$ is always nonsingular with inverse

$$\exp(A)^{-1} \;=\; \exp(-A).$$

In general the expansion of $\exp(A+B)$ is somewhat more complicated. From the following result only the very first inequality is needed later, but the full result may be of interest for curious readers and illustrates why treating $L_\rho(\Sigma, Q)$ as a function of $\log(\Sigma)$ is not that straightforward.

Lemma 5.1 (Taylor expansions of $\exp(\cdot)$)**.** *For matrices $A, \Delta \in \mathbb{R}^{q\times q}$,*

$$\begin{aligned}\exp(A+\Delta) \;&=\; \exp(A) + R_0(A,\Delta)\\ &=\; \exp(A) + \int_0^1 \exp((1-u)A)\Delta\exp(uA)\,du + R_1(A,\Delta)\end{aligned}$$

with

$$\|R_0(A,\Delta)\| \;\le\; e^{\|A\|+\|\Delta\|}\|\Delta\| \quad \textit{and} \quad \|R_1(A,\Delta)\| \;\le\; e^{\|A\|+\|\Delta\|}\|\Delta\|^2/2.$$

Moreover,

$$\exp(A+\Delta) \;=\; \exp(A) + \sum_{k=1}^{\infty} \frac{1}{k!}\,\mathbb{E}\big[\exp(U_{k0}A)\Delta\exp(U_{k1}A)\cdots\Delta\exp(U_{kk}A)\big],$$

where $U_{k0} = 1 - \sum_{j=1}^k U_{kj}$, and $(U_{kj})_{j=1}^k$ is uniformly distributed on the convex polytope $\big\{u \in [0,1]^k : \sum_{j=1}^k u_j \le 1\big\}$.

The exponential transform on $\mathbb{R}^{q\times q}_{\mathrm{sym}}$ Any matrix $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$ may be written as

$$A \;=\; \sum_{i=1}^{q} \lambda_i(A)\,u_i u_i^\top \;=\; U\,\mathrm{diag}\big((\lambda_i(A))_{i=1}^q\big)U^\top$$

with the ordered eigenvalues $\lambda_i(A)$ of A, an orthonormal basis $u_1, u_2, \ldots, u_q$ of corresponding eigenvectors, and the orthogonal matrix $U = [u_1\, u_2\, \ldots\, u_q]$. Then one can easily verify that

$$\exp(A) \;=\; \sum_{i=1}^{q} \exp(\lambda_i(A))\, u_i u_i^\top \;\in\; \mathbb{R}^{q\times q}_{\mathrm{sym},>0}.$$

As a mapping from $\mathbb{R}^{q\times q}_{\mathrm{sym}}$ to $\mathbb{R}^{q\times q}_{\mathrm{sym},>0}$, the exponential function is bijective with inverse

$$\log(A) \;:=\; \sum_{i=1}^{q} \log(\lambda_i(A))\, u_i u_i^\top.$$

Moreover,

$$\det(\exp(A)) \;=\; \exp(\mathrm{tr}(A)).$$

Local parametrizations of $\mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ Unfortunately, for $\Sigma, \Sigma' \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$, the equation $\Sigma' = \exp(\log(\Sigma) + A)$ with $A := \log(\Sigma') - \log(\Sigma)$ is not very helpful, because the Taylor expansion of $\exp(\log(\Sigma) + A)$ is somewhat awkward, unless $\log(\Sigma)$ and A are interchangeable. In view of our considerations on linear equivariance, we consider a different approach: Let $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$, and fix an arbitrary $B \in \mathbb{R}^{q\times q}_{\mathrm{ns}}$ such that

$$\Sigma \;=\; BB^\top,$$

e.g. $B = \Sigma^{1/2}$. Then

$$\mathbb{R}^{q\times q}_{\mathrm{sym},>0} \;=\; \big\{B\exp(A)B^\top : A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}\big\}.$$

Indeed, any matrix $\Sigma' \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ may be written as $B\exp(A)B^\top$ with $A := \log(B^{-1}\Sigma' B^{-\top})$. Note that the matrix A depends on both B and Σ', but its eigenvalues are simply $\lambda_i(A) = \log\lambda_i(\Sigma^{-1}\Sigma')$. Moreover, if $\det(\Sigma) = 1$, then

$$\big\{\Sigma' \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0} : \det(\Sigma') = 1\big\} \;=\; \big\{B\exp(A)B^\top : A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}, \mathrm{tr}(A) = 0\big\}.$$

5.2. First-order smoothness of the criterion function

We start with an expansion of $L_\rho(\cdot, Q)$ in small neighborhoods of I_q. To this end we need the matrix

$$\Psi_\rho(Q) \;:=\; \Psi_\rho(I_q, Q) = \int \rho'(\mathrm{tr}(M))M\, Q(dM) \;\in\; \mathbb{R}^{q\times q}_{\mathrm{sym}}.$$

Proposition 5.2 (1st order Taylor expansion)**.** *For $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$,*

$$L_\rho(\exp(A), Q) \;=\; \langle A, G_\rho(Q)\rangle + R_\rho(A, Q)$$

with the gradient

$$G_\rho(Q) \;:=\; I_q - \Psi_\rho(Q) \;\in\; \mathbb{R}^{q\times q}_{\mathrm{sym}}$$

and a remainder $R_\rho(A,Q)$ satisfying the following inequalities:

$$\begin{aligned}\big|\langle A, G_\rho(Q)\rangle\big| &\le (q + J_\rho(Q))\|A\|,\\ |R_\rho(A,Q)| &\le \big(J_\rho(e^{\|A\|},Q) - J_\rho(e^{-\|A\|},Q)\big)\|A\| + J_\rho(Q)\|A\|^2/2,\end{aligned}$$

where $J_\rho(Q) := J_\rho(1,Q)$ and

$$J_\rho(\lambda, Q) := \int \psi(\lambda \operatorname{tr}(M))\, Q(dM).$$

Note that $J_\rho(\cdot,Q) \equiv q$ in Case 0. In Case 1, $J_\rho(\lambda,Q)$ is continuous and monotone increasing in $\lambda > 0$ with values in $[0, \psi(\infty))$. Thus in both cases,

$$|R_\rho(A,Q)| = o(\|A\|) \quad \text{as } A \to 0.$$

Proposition 5.2 carries over to expansions in other neighborhoods via linear equivariance: For any fixed $B \in \mathbb{R}^{q\times q}_{\mathrm{ns}}$ and $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$ we have by (4.4),

$$\begin{aligned}L_\rho(B\exp(A)B^\top, Q) - L_\rho(BB^\top, Q) &= L_\rho(\exp(A), Q_B)\\ &= \langle A, G_\rho(Q_B)\rangle + R_\rho(A, Q_B),\end{aligned}$$

where

$$\begin{aligned}\big|\langle A, G_\rho(Q_B)\rangle\big| &\le (q + J_\rho(Q_B))\|A\|,\\ |R_\rho(A,Q_B)| &\le \big(J_\rho(e^{\|A\|},Q_B) - J_\rho(e^{-\|A\|},Q_B)\big)\|A\| + J_\rho(Q_B)e^{\|A\|}\|A\|^2/2.\end{aligned}$$

Moreover, with $\Sigma := BB^\top$, Lemma 4.7 and monotonicity of ψ yield

$$J_\rho(\lambda, Q_B) = \int \psi(\lambda \operatorname{tr}(\Sigma^{-1}M))\, Q(dM) \le J_\rho(\lambda/\lambda_{\min}(\Sigma), Q). \tag{5.1}$$

Note also that

$$G_\rho(Q_B) = B^{-1}\big(\Sigma - \Psi_\rho(\Sigma, Q)\big)B^{-\top},$$

so the fixed-point equation (4.7) in Theorem 4.9 is satisfied if, and only if, $G_\rho(Q_B) = 0$.

Proposition 5.2 implies also that $L_\rho(\cdot,Q)$ is a continuously differentiable and locally Lipschitz-continuous function on $\mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ in the usual sense:

Corollary 5.3 (Smoothness)**.** *The function $L_\rho(\cdot,Q)$ is continuously differentiable on $\mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ with gradient*

$$\begin{aligned}\nabla L_\rho(\Sigma, Q) &= \Sigma^{-1} - \int \rho'(\operatorname{tr}(\Sigma^{-1}M))\Sigma^{-1}M\Sigma^{-1}\, Q(dM)\\ &= B^{-1}G_\rho(Q_B)B^{-1}\end{aligned}$$

with $B := \Sigma^{1/2}$. Moreover, let K be a convex subset of $\mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ with $\lambda_{\min}(K) := \inf_{\Sigma\in K}\lambda_{\min}(\Sigma) > 0$. Then for $\Sigma_0, \Sigma_1 \in K$,

$$\big|L_\rho(\Sigma_1, Q) - L_\rho(\Sigma_0, Q)\big| \le \big(q + J_\rho(\lambda_{\min}(K)^{-1}, Q)\big)\lambda_{\min}(K)^{-1}\|\Sigma_1 - \Sigma_0\|.$$

5.3. Convexity and coercivity

Theorem 4.9 follows essentially from the next two results. The first one provides a surrogate for the simpler claim that $L_\rho(\Sigma, Q)$ is a convex function of $\log(\Sigma)$. The second one deals with the behavior of $L_\rho(\Sigma, Q)$ as $\|\log(\Sigma)\| \to \infty$.

Proposition 5.4 (Convexity)**.** *For any fixed $B \in \mathbb{R}^{q\times q}_{\mathrm{ns}}$ and $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$,*

$$\mathbb{R} \ni t \;\mapsto\; L_\rho(B\exp(tA)B^\top, Q)$$

is a convex function. This convexity is strict if, and only if,

$$\begin{cases} Q\big(\bigcup_{i=1}^{\ell} \mathbb{M}(B\mathbb{V}_i)\big) \;<\; 1 & \text{in Case 0,} \\ Q\big(\mathbb{M}(B\mathbb{V}_0)\big) \;<\; 1 & \text{in Case 1,} \end{cases}$$

where $\mathbb{V}_1, \ldots, \mathbb{V}_\ell$ are the eigenspaces of A, and $\mathbb{V}_0 := \{x \in \mathbb{R}^q : Ax = 0\}$.

Proposition 5.5 (Coercivity)**.** *Let B be an arbitrary fixed matrix in $\mathbb{R}^{q\times q}_{\mathrm{ns}}$. In Case 0,*

$$\lim_{\|A\|\to\infty,\, \mathrm{tr}(A)=0} L_\rho(B\exp(A)B^\top, Q) \;=\; \infty$$

if, and only if, Condition 0 is true. In Case 1,

$$\lim_{\|A\|\to\infty} L_\rho(B\exp(A)B^\top, Q) \;=\; \infty$$

if, and only if, Condition 1 is true.

The convexity property in Proposition 5.4 is sometimes called "geodesic convexity" (cf. Wiesel, 2012). This name stems from the fact that for arbitrary matrices $\Sigma_0 = BB^\top$ and $\Sigma_1 = B\exp(A)B^\top$ in $\mathbb{R}^{q\times q}_{\mathrm{sym},>0}$, the path

$$[0,1] \ni t \;\mapsto\; \Gamma(t) := B\exp(tA)B^\top$$

minimizes the "length"

$$\int_0^1 \big\|\Gamma(t)^{-1/2}\Gamma'(t)\Gamma(t)^{-1/2}\big\|_F \, dt$$

over all continuously differentiable functions $\Gamma : [0,1] \to \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ with $\Gamma(0) = \Sigma_0$ and $\Gamma(1) = \Sigma_1$; see Bhatia (2007, Chapter 6).

5.4. Second-order smoothness of the criterion function

In order to prove differentiability of $\mathbf{\Sigma}_\rho(\cdot)$, we need second order Taylor expansions of $L_\rho(\cdot, Q)$. These are also useful to replace the fixed-point algorithm described earlier by faster methods, see Dümbgen et al. (2013).

From now on we assume that ρ is twice continuously differentiable on $(0,\infty)$. In addition to $\psi(s) = s\rho'(s)$ we consider

$$\psi_2(s) \;:=\; s\psi'(s) \;=\; \psi(s) + s^2\rho''(s).$$

In Case 0, $\psi \equiv q$, so $\psi' \equiv \psi_2 \equiv 0$. Case 1 is modified as follows:

Case 1' We assume that $\psi' > 0$ and that ψ has limits $\psi(0) = 0$ and $\psi(\infty) \in (q, \infty]$. Moreover we assume that

$$\int \psi(\mathrm{tr}(M))\, Q(dM) \;<\; \infty \tag{5.2}$$

and that there exists a constant $\kappa > 0$ such that

$$\psi_2(s) \;\le\; \kappa\psi(s) \quad \text{for all } s > 0. \tag{5.3}$$

Remark 5.6. Inequality (5.3) is mainly for convenience and to avoid additional integrability conditions for ψ_2. It also allows to replace (4.3) with the simpler condition (5.2), see Lemma 5.10 below. It follows from $\psi_2(s) = \psi(s) + s^2\rho''(s)$ and $\psi, \psi' > 0$ that $s^2\rho''(s) = \psi_2(s) - \psi(s) \in \big(-\psi(s), \psi_2(s)\big)$. Hence inequality (5.3) is equivalent to the existence of a constant $\tilde\kappa$ such that

$$s^2|\rho''(s)| \;\le\; \tilde\kappa\psi(s) \quad \text{for all } s > 0. \tag{5.4}$$

Remark 5.7. Suppose that ρ' is non-increasing, i.e. $\rho'' \le 0$. Then $0 < \psi_2(s) \le \psi(s)$ and $-\psi(s) < s^2\rho''(s) \le 0$. Hence (5.3) and (5.4) are satisfied with $\kappa = \tilde\kappa = 1$.

Example 5.8 (Multivariate elliptical Weibull-distributions)**.** In case of $\rho(s) := s^\gamma$ for a constant $\gamma > 0$, we have $\psi(s) = \gamma s^\gamma$ and

$$s^2\rho''(s) \;=\; (\gamma - 1)\psi(s), \quad \psi_2(s) \;=\; \gamma\psi(s),$$

so (5.3) and (5.4) are satisfied with $\kappa = \gamma$ and $\tilde\kappa = |\gamma - 1|$.

Example 5.9. In case of $\rho(s) := (\nu + q)\log(1 + s^2)/2$ for a constant $\nu > 0$, we have $\psi(s) = (\nu + q)s^2/(1 + s^2)$ and

$$s^2\rho''(s) \;=\; \big(1 - 2\psi(s)/\psi(\infty)\big)\psi(s), \quad \psi_2(s) \;=\; 2\big(1 - \psi(s)/\psi(\infty)\big)\psi(s),$$

so (5.3) and (5.4) are satisfied with $\kappa = 2$ and $\tilde\kappa = 1$.

Lemma 5.10. *Let $\phi : (0, \infty) \to (0, \infty)$ be a differentiable function. For any $\kappa \in \mathbb{R}$ the following two statements are equivalent:*

$$s\phi'(s) \;\le\; \kappa\phi(s) \quad \textit{for all } s > 0; \tag{5.5}$$
$$\phi(\lambda s) \;\le\; \lambda^\kappa\phi(s) \quad \textit{for all } s > 0 \textit{ and } \lambda > 1. \tag{5.6}$$

Now we are ready to extend the expansion of $L_\rho(\cdot, Q)$ around I_q from Proposition 5.2:

Proposition 5.11 (2nd order Taylor expansion)**.** *In Case 0 and Case 1', for arbitrary $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$,*

$$L_\rho(\exp(A), Q) \;=\; \langle A, G_\rho(Q)\rangle + 2^{-1}H_\rho(A, Q) + R_{\rho,2}(A, Q) \tag{5.7}$$

with the gradient $G_\rho(Q)$ as in Proposition 5.2, the quadratic term

$$\begin{aligned} H_\rho(A,Q) &:= \int \big(\rho'(\mathrm{tr}(M))\,\mathrm{tr}(A^2M) + \rho''(\mathrm{tr}(M))\,\mathrm{tr}(AM)^2\big)\,Q(dM) \\ &= \mathrm{tr}(A^2\Psi_\rho(Q)) + \int \rho''(\mathrm{tr}(M))\,\mathrm{tr}(AM)^2\,Q(dM) \end{aligned}$$

and a remainder term $R_{\rho,2}(A,Q)$ satisfying the following inequalities:

$$H_\rho(A,Q) \in \big[0,(1+\kappa)J_\rho(Q)\|A\|^2\big], \tag{5.8}$$

$$|R_{\rho,2}(A,Q)| \le \Omega(\|A\|,Q)\|A\|^2/2 + (\kappa+1/7)J_\rho(Q)\|A\|^3 \tag{5.9}$$

with

$$\Omega(\delta,Q) := \int \sup_{z\in[-\delta,\delta]} \big|\psi_2(e^z\,\mathrm{tr}(M)) - \psi_2(\mathrm{tr}(M))\big|\,Q(dM).$$

Moreover,

$$H_\rho(A,Q) > 0 \quad \text{if} \begin{cases} Q\big(\bigcup_{i=1}^\ell \mathbb{M}(\mathbb{V}_i)\big) < 1 & \text{in Case 0,} \\ Q(\mathbb{M}(\mathbb{V}_0)) < 1 & \text{in Case 1',} \end{cases} \tag{5.10}$$

where $\mathbb{V}_1,\dots,\mathbb{V}_\ell$ are the eigenspaces of A, and $\mathbb{V}_0 := \{x\in\mathbb{R}^q : Ax = 0\}$.

Note that $\Omega(\delta,Q)$ is continuous in $\delta\ge 0$ with $\Omega(0,Q)=0$. This follows from the fact that

$$\sup_{z\in[-\delta,\delta]} \big|\psi_2(e^z\,\mathrm{tr}(M)) - \psi_2(\mathrm{tr}(M))\big|$$

is continuous in $\delta\ge 0$ and not greater than $\kappa\psi(e^\delta\,\mathrm{tr}(M)) \le \kappa e^{\kappa\delta}\psi(\mathrm{tr}(M))$. In particular,

$$R_{\rho,2}(A,Q) = o(\|A\|^2) \quad \text{as } A\to 0.$$

Again Proposition 5.11 carries over to expansions in other neighborhoods via linear equivariance: For any fixed $B\in\mathbb{R}^{q\times q}_{\mathrm{ns}}$ and $A\in\mathbb{R}^{q\times q}_{\mathrm{sym}}$,

$$\begin{aligned} &L_\rho(B\exp(A)B^\top,Q) - L_\rho(BB^\top,Q) \\ &\quad = L_\rho(\exp(A),Q_B) = \langle A, G_\rho(Q_B)\rangle + 2^{-1}H_\rho(A,Q_B) + R_{\rho,2}(A,Q_B), \end{aligned}$$

where $R_{\rho,2}(A,Q_B) = o(\|A\|^2)$ as $A\to 0$.

The Hessian operator The quadratic term $H_\rho(A,Q)$ in Proposition 5.11 may be written as

$$H_\rho(A,Q) = \langle A, H_\rho(Q)A\rangle$$

with the linear operator $H_\rho(Q):\mathbb{R}^{q\times q}_{\mathrm{sym}}\to\mathbb{R}^{q\times q}_{\mathrm{sym}}$ given by

$$\begin{aligned} H_\rho(Q)A &:= \int \big(\rho'(\mathrm{tr}(M))2^{-1}(AM+MA) + \rho''(\mathrm{tr}(M))\,\mathrm{tr}(AM)M\big)\,Q(dM) \\ &= 2^{-1}\big(A\Psi_\rho(Q) + \Psi_\rho(Q)A\big) + \int \rho''(\mathrm{tr}(M))\,\mathrm{tr}(AM)M\,Q(dM). \end{aligned}$$

This operator is self-adjoint, that means, $\langle A, H_\rho(Q)B\rangle = \langle B, H_\rho(Q)A\rangle$ for arbitrary $A,B\in\mathbb{R}^{q\times q}_{\mathrm{sym}}$.

Invertibility in Case 1' Under Condition 1 it follows from the last part of Proposition 5.11 that $H_\rho(Q)$ is positive definite and thus invertible.

Invertibility in Case 0 The gradient $G_\rho(Q) = I_q - q\int \mathrm{tr}(M)^{-1} M\, Q(dM)$ is contained in the linear subspace

$$\mathbb{W}_0 \;:=\; \{A \in \mathbb{R}^{q\times q}_{\mathrm{sym}} : \mathrm{tr}(A) = 0\},$$

and for any $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$,

$$H_\rho(Q)A \;=\; q\int \big(\mathrm{tr}(M)^{-1}2^{-1}(AM + MA) - \mathrm{tr}(M)^{-2}\,\mathrm{tr}(AM)M\big)\, Q(dM)$$

belongs to $\mathbb{W}_0$, too. Hence we view $H_\rho(Q)$ as a linear operator from $\mathbb{W}_0$ to $\mathbb{W}_0$. Under Condition 0, the last part of Proposition 5.11 implies that this operator is positive definite und thus invertible.

6. Continuity, consistency and differentiability

In this section we derive various properties of $\boldsymbol{\Sigma}_\rho(\cdot)$ and related limit theorems. The arguments we use are adaptations of standard arguments in the statistical literature, e.g. the monographs mentioned in the introduction. Related are also the papers by Haberman (1989) and Niemiro (1992) about M-estimation with convex criterion functions.

Throughout this section let Q be a distribution in $\mathcal{Q}_\rho$ and define

$$\mathbb{Y} \;:=\; \begin{cases} \mathbb{R}^{q\times q}_{\mathrm{sym},\geq 0} \setminus \{0\} & \text{in Case 0,} \\ \mathbb{R}^{q\times q}_{\mathrm{sym},\geq 0} & \text{in Case 1.} \end{cases}$$

Moreover we consider the linear space

$$\mathbb{W} \;:=\; \begin{cases} \{A \in \mathbb{R}^{q\times q}_{\mathrm{sym}} : \mathrm{tr}(A) = 0\} & \text{in Case 0,} \\ \mathbb{R}^{q\times q}_{\mathrm{sym}} & \text{in Case 1.} \end{cases}$$

Recall that in Case 1', $H_\rho(Q) : \mathbb{W} \to \mathbb{W}$ is an invertible linear operator.

Unless stated otherwise, all subsequent asymptotic statements refer to the sequence index n tending to ∞. Furthermore, "$\to_p$" and "$\to_w$" stand for convergence in probability and weak convergence, respectively.

6.1. Continuity

Our first result establishes a certain continuity property of $\boldsymbol{\Sigma}_\rho(\cdot)$.

Theorem 6.1 (Continuity I)**.** *Let $(Q_n)_n$ be a sequence of probability distributions on $\mathbb{Y}$ converging weakly to Q. In Case 1 suppose in addition that all Q_n satisfy (4.3) and that*

$$\int \psi\big(\lambda_o\,\mathrm{tr}(\boldsymbol{\Sigma}_\rho(Q)^{-1}M)\big)\, Q_n(dM) \;\to\; \int \psi\big(\lambda_o\,\mathrm{tr}(\boldsymbol{\Sigma}_\rho(Q)^{-1}M)\big)\, Q(dM) \quad (6.1)$$

for some $\lambda_o > 1$. Then $Q_n \in \mathcal{Q}_\rho$ for sufficiently large n, and

$$\mathbf{\Sigma}_\rho(Q_n) \;\to\; \mathbf{\Sigma}(Q).$$

Remark 6.2 (Weak Continuity)**.** In case of $\psi(\infty) < \infty$, Condition (6.1) is satisfied for any $\Sigma_o \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ because $Q_n \to_w Q$. Thus Theorem 6.1 shows that the set $\mathcal{Q}_\rho$ is open in the topology of weak convergence of probability measures on $\mathbb{Y}$, and that the functional $\mathbf{\Sigma}_\rho$ is weakly continuous on $\mathcal{Q}_\rho$.

Our proof of Theorem 6.1 covers also the situation of random distributions $\widehat{Q}_n$ in place of Q_n. Indeed the following result is true:

Theorem 6.3 (Continuity II)**.** *Let $\widehat{Q}_1, \widehat{Q}_2, \widehat{Q}_3, \ldots$ be random distributions on $\mathbb{Y}$ such that for any bounded and continuous function $f : \mathbb{Y} \to \mathbb{R}$,*

$$\int f \, d\widehat{Q}_n \;\to_p\; \int f \, dQ. \tag{6.2}$$

In Case 1 suppose further that $\widehat{Q}_n$ satisfies (4.3) almost surely and that

$$\int \psi\big(\lambda_o \operatorname{tr}(\mathbf{\Sigma}_\rho(Q)^{-1}M)\big)\, \widehat{Q}_n(dM) \;\to_p\; \int \psi\big(\lambda_o \operatorname{tr}(\mathbf{\Sigma}_\rho(Q)^{-1}M)\big)\, Q(dM) \tag{6.3}$$

for some $\lambda_o > 1$. Then $\mathbb{P}(\widehat{Q}_n \in \mathcal{Q}_\rho) \to 1$ and

$$\mathbf{\Sigma}_\rho(\widehat{Q}_n) \;\to_p\; \mathbf{\Sigma}_\rho(Q).$$

In case of $\psi(\infty) < \infty$, one could derive Theorem 6.3 easily from Theorem 6.1 by means of metrics for weak convergence as described in by Dudley (2002, Section 11.3). In the general setting, however, it is easier to prove Theorem 6.3 directly and realize that Theorem 6.1 is just a special case of it.

6.2. Differentiability

In this subsection we refine Theorem 6.3 with an asymptotic linear expansion of $\mathbf{\Sigma}_\rho(\cdot)$ in Cases 0 and 1'. By linear equivariance it suffices to consider the case

$$\mathbf{\Sigma}_\rho(Q) \;=\; I_q.$$

Theorem 6.4 (Differentiability)**.** *Let $\widehat{Q}_1$, $\widehat{Q}_2$, $\widehat{Q}_3$, ... be random distributions on $\mathbb{Y}$ satisfying Condition (6.2). In Case 1' suppose further that for all n, $\int \psi(\operatorname{tr}(M))\, \widehat{Q}_n(dM) < \infty$ almost surely, and*

$$\int \psi(\operatorname{tr}(M))\, \widehat{Q}_n(dM) \;\to_p\; \int \psi(\operatorname{tr}(M))\, Q(dM). \tag{6.4}$$

Then in Cases 0 and 1',

$$G_\rho(\widehat{Q}_n) \;\to_p\; 0, \quad \text{and} \quad H_\rho(\widehat{Q}_n) \;\to_p\; H_\rho(Q).$$

Moreover, $\mathbb{P}(\widehat{Q}_n \in \mathcal{Q}_\rho) \to 1$ and

$$\log(\mathbf{\Sigma}_\rho(\widehat{Q}_n)) \;=\; -H_\rho(Q)^{-1} G_\rho(\widehat{Q}_n) + o_p\big(\|G_\rho(\widehat{Q}_n)\|\big). \tag{6.5}$$

Remark 6.5. Condition (6.4) seems to be weaker than (6.3) at first glance. But in Case 1',

$$\psi\big(\lambda_o \operatorname{tr}(M)\big) \;\le\; \lambda_o^\kappa \psi(\operatorname{tr}(M))$$

for any $\lambda_o > 1$ and $M \in \mathbb{Y}$ by Lemma 5.10. Consequently (6.3) follows from (6.2) and (6.4) by virtue of Lemma 8.5 in Section 8.

Remark 6.6. Note that the asymptotic expansion (6.5) is equivalent to the expansion

$$\boldsymbol{\Sigma}_\rho(\widehat{Q}_n) \;=\; I_q - H_\rho(Q)^{-1} G_\rho(\widehat{Q}_n) + o_p\big(\|G_\rho(\widehat{Q}_n)\|\big).$$

Remark 6.7 (Weak Differentiability)**.** In Cases 0 and 1' with $\psi(\infty) < \infty$, Theorem 6.4 shows that the functional $\boldsymbol{\Sigma}_\rho$ is weakly differentiable on $\mathcal{Q}_\rho$ in the following sense: Let $Q \in \mathcal{Q}_\rho$ and $B := \boldsymbol{\Sigma}_\rho(Q)^{1/2}$. Further let $(Q_n)_n$ be a sequence of probability distributions in $\mathcal{Q}_\rho$ converging weakly to Q. Then $G_\rho((Q_n)_B) \to 0$ and

$$\log(B^{-1}\boldsymbol{\Sigma}_\rho(Q_n)B^{-1}) \;=\; -H_\rho(Q_B)^{-1} G_\rho((Q_n)_B) + o\big(\big\|G_\rho((Q_n)_B)\big\|\big).$$

6.3. Orthogonally invariant distributions

The previous differentiability results involve the operator $H_\rho(Q)$. The latter turns out to have a special structure under a certain symmetry condition on Q:

Definition 6.8 (Orthogonal symmetry)**.** The distribution Q of a random matrix $M \in \mathbb{R}^{q\times q}_{\mathrm{sym},\ge 0}$ is called orthogonally invariant if

$$\mathcal{L}(VMV^\top) \;=\; \mathcal{L}(M) \quad \text{for any orthogonal matrix } V \in \mathbb{R}^{q\times q}.$$

This property is closely related to spherically symmetric distributions on $\mathbb{R}^q$. For instance, let $Q = \mathcal{L}(XX^\top)$ with a random vector X with spherically symmetric distribution on $\mathbb{R}^q$. Then Q is orthogonally invariant. Another example is given by $Q = \mathcal{L}(S(X_1, X_2, \ldots, X_k))$ with independent, identically distributed random vectors $X_1, X_2, \ldots, X_k \in \mathbb{R}^q$ such that $\mathcal{L}(X_1 - \mu)$ is spherically symmetric for some $\mu \in \mathbb{R}^q$.

By linear equivariance of $\boldsymbol{\Sigma}_\rho(\cdot)$, orthogonal invariance of Q implies that $\boldsymbol{\Sigma}_\rho(Q)$ is a positive multiple of I_q. As shown in the subsequent lemma, the operator $H_\rho(Q)$ has a rather simple form here. It will be convenient to decompose $\mathbb{R}^{q\times q}_{\mathrm{sym}}$ as

$$\mathbb{R}^{q\times q}_{\mathrm{sym}} \;=\; \mathbb{W}_0 + \mathbb{W}_1$$

with $\mathbb{W}_0 = \{A \in \mathbb{R}^{q\times q}_{\mathrm{sym}} : \operatorname{tr}(A) = 0\}$ and $\mathbb{W}_1 := \{sI_q : s \in \mathbb{R}\}$. Any matrix $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$ has the unique decomposition

$$A \;=\; A_0 + A_1$$

with $A_0 := A - q^{-1}\operatorname{tr}(A) I_q \in \mathbb{W}_0$ and $A_1 := q^{-1}\operatorname{tr}(A) I_q \in \mathbb{W}_1$.

Lemma 6.9. *Suppose that Q is orthogonally invariant, and let $\mathbf{\Sigma}_\rho(Q) = I_q$. Then for $A = A_0 + A_1$ with $A_0 \in \mathbb{W}_0, A_1 \in \mathbb{W}_1$,*

$$H_\rho(Q)A \;=\; d_0(Q)A_0 + d_1(Q)A_1,$$

where

$$\begin{aligned} d_0(Q) \;&:=\; 1 + \frac{2}{q(q+2)} \int \rho''(\mathrm{tr}(M)) \Big(\|M\|_F^2 + \frac{\mathrm{tr}(M)^2 - \|M\|_F^2}{q-1} \Big)\, Q(dM), \\ d_1(Q) \;&:=\; 1 + \frac{1}{q} \int \rho''(\mathrm{tr}(M))\, \mathrm{tr}(M)^2\, Q(dM). \end{aligned}$$

Implications for rank one distributions Suppose that a random matrix $M \sim Q$ satisfies $\mathrm{rank}(M) \le 1$ almost surely. This is true in settings (4.1) and (4.2) with $k = 2$. Then $\|M\|_F = \mathrm{tr}(M)$ almost surely, so

$$\begin{aligned} d_0(Q) \;&=\; 1 + \frac{2}{q(q+2)} \int \rho''(\mathrm{tr}(M))\, \mathrm{tr}(M)^2\, Q(dM), \\ d_1(Q) \;&=\; 1 + \frac{1}{q} \int \rho''(\mathrm{tr}(M))\, \mathrm{tr}(M)^2\, Q(dM). \end{aligned}$$

Implications for Case 0 Recall that in Case 0, $\rho(s) = q\log(s)$, so $\rho'(s) = q/s$ and $\rho''(s) = -q/s^2$. Thus $d_1(Q) = 0$, and for $A = A_0 + A_1$ with $A_0 \in \mathbb{V}_0, A_1 \in \mathbb{V}_1$,

$$H_\rho(Q)A \;=\; d_0(Q)A_0$$

with

$$d_0(Q) \;=\; 1 - \frac{2}{q+2} \int \frac{(q-2)\|M\|_F^2/\,\mathrm{tr}(M)^2 + 1}{q-1}\, Q(dM).$$

In particular, if $\mathrm{rank}(M) = 1$ almost surely, then

$$H_\rho(Q)A \;=\; \frac{q}{q+2}\, A_0.$$

6.4. Consistency and Central Limit Theorems

In this section we apply the previous results to particular empirical distributions related to Settings (4.1) and (4.2). For convenience we restrict our attention to Cases 0 and 1'.

For some fixed integer $k \ge 1$ and arbitrary integers $n \ge k$ we consider distributions

$$Q \;:=\; Q^k(P) \quad \text{and} \quad Q_n \;:=\; Q^k(P_n)$$

in $\mathcal{Q}_\rho$ with distributions P, P_n on $\mathbb{R}^q$ such that

$$\mathbf{\Sigma}_\rho(Q) \;=\; I_q \;=\; \mathbf{\Sigma}_\rho(Q_n) \quad \text{for all } n \ge k.$$

Recall that in Case 0, $\tilde{Q} = Q^k(\tilde{P}) \in \mathcal{Q}_\rho$ implies that

$$\begin{cases} \tilde{P}(\{0\}) = 0 & \text{if } k = 1, \\ \tilde{P}(\{x\}) = 0 \;\text{ for all } x \in \mathbb{R}^q & \text{if } k \ge 2. \end{cases}$$

Additional assumptions We assume that

$$P_n \;\to_w\; P.$$

Further, for a certain exponent $m \geq 1$ we assume that

$$\int \psi(\|x\|^2)^m \, P_n(dx) \;\to\; \int \psi(\|x\|^2)^m \, P_n(dx),$$

where all integrals on the left and right hand side are finite.

Note that for any exponent $m \geq 1$, the second part of the additional assumptions is a consequence of the first part whenever $\psi(\infty) < \infty$.

Now we consider for $n \geq k$ independent random vectors $X_{n1}, X_{n2}, \ldots, X_{nn}$ with distribution P_n and define

$$\widehat{Q}_n \;:=\; \begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^n \delta_{X_{ni}X_{ni}^\top} & \text{if } k = 1, \\ \dbinom{n}{k}^{-1} \displaystyle\sum_{1 \leq i_1 < \cdots < i_k \leq n} \delta_{S(X_{ni_1}, \ldots, X_{ni_k})} & \text{if } k \geq 2. \end{cases}$$

Our first result proves consistency of $\mathbf{\Sigma}_\rho(\widehat{Q}_n)$ as an estimator for $\mathbf{\Sigma}_\rho(Q_n) = I_q$. It is essentially a corollary to Theorem 6.3:

Theorem 6.10 (Consistency)**.** *In the setting just described, suppose that the additional assumptions hold with $m = 1$. Then $\mathbb{P}(\widehat{Q}_n \in \mathcal{Q}_\rho) \to 1$ and*

$$\mathbf{\Sigma}_\rho(\widehat{Q}_n) \;\to_p\; I_q.$$

Our second result provides a precise linear expansion for $\mathbf{\Sigma}_\rho(\widehat{Q}_n)$ and is based on Theorem 6.4:

Theorem 6.11 (Linear expansion)**.** *Let $\mathbb{X} := \mathbb{R}^q \setminus \{0\}$ in Case 0 with $k = 1$, and $\mathbb{X} := \mathbb{R}^q$ otherwise. In the just described setting, suppose that the additional assumptions hold with $m = 2$. Then $\mathbb{P}(\widehat{Q}_n \in \mathcal{Q}_\rho) \to 1$ and*

$$\sqrt{n}\log(\mathbf{\Sigma}_\rho(\widehat{Q}_n)) \;=\; \frac{1}{\sqrt{n}}\sum_{i=1}^n \big(Z(X_{ni}) - \mathbb{E}Z(X_{n1})\big) + o_p(1)$$

for some continuous function $Z : \mathbb{X} \to \mathbb{R}^{q\times q}_{\mathrm{sym}}$ depending only on P such that

$$\sup_{x\in\mathbb{X}} \frac{\|Z(x)\|}{1 + \psi(\|x\|^2)} \;<\; \infty \quad \text{and} \quad \int Z \, dP \;=\; 0.$$

Precisely, if $k = 1$, then

$$Z(x) \;:=\; H_\rho(Q)^{-1}\big(\rho'(\|x\|^2)xx^\top - I_q\big) \quad \text{and} \quad \mathbb{E}Z(X_{n1}) \;=\; 0.$$

If $k \geq 2$, then

$$Z(x) \;=\; kH_\rho(Q)^{-1}\Big(\mathbb{E}\big[\rho'\big(\mathrm{tr}(S(x, X_2, \ldots, X_k))\big)S(x, X_2, \ldots, X_k)\big] - I_q\Big)$$

with independent random vectors $X_2, \ldots, X_k \sim P$.

Remark 6.12 (Central Limit Theorem)**.** By virtue of the multivariate version of Lindeberg's Central Limit Theorem, the expansion in Theorem 6.11 implies a Central Limit Theorem for the estimator $\mathbf{\Sigma}_\rho(\widehat{Q}_n)$. Namely,

$$\mathcal{L}\big(\sqrt{n}\log(\mathbf{\Sigma}_\rho(\widehat{Q}_n))\big) \;\to_w\; \mathcal{N}_{q\times q}\big(0, \mathrm{Cov}(Z(X))\big)$$

with $X \sim P$. This means, that for any matrix $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$,

$$\big\langle \sqrt{n}\log(\mathbf{\Sigma}_\rho(\widehat{Q}_n)), A\big\rangle \;\to_w\; \mathcal{N}\big(0, \mathrm{Var}(\langle Z(X), A\rangle)\big).$$

Remark 6.13 (Spherical symmetry I)**.** Let P be spherically symmetric around $0 \in \mathbb{R}^q$. Then the matrix-valued function Z in Theorem 6.11 may be written as

$$Z(x) \;=\; z_0(\|x\|^2)xx^\top + z_1(\|x\|^2)I_q$$

with certain functions $z_0, z_1 : [0,\infty) \to \mathbb{R}$, where $z_1(s) = -q^{-1}sz_0(s)$ in Case 0.

Remark 6.14 (Spherical symmetry II)**.** Let P be spherically symmetric around $0 \in \mathbb{R}^q$, and let $k = 1$. Further let

$$\rho(s) \;=\; (\nu+q)\log(\nu+s)$$

with $\nu = 0$ (Case 0) or $\nu > 0$ (Case 1'). For $x \in \mathbb{R}^q$ we write

$$xx^\top \;=\; A_0(x) + a(x)I_q + I_q$$

with $a(x) := q^{-1}\|x\|^2 - 1$, so that $\mathrm{tr}(A_0(x)) = 0$. Then the matrix-valued function Z in Theorem 6.11 is given by

$$Z(x) \;=\; (\nu+\|x\|^2)^{-1}\big(c_0A_0(x) + c_1a(x)I_q\big)$$

with

$$c_0 \;:=\; \frac{(q+\nu)(q+2)}{q+2(1-\beta)\nu/q}, \quad c_1 \;:=\; 1_{[\nu>0]}\frac{q}{1-\beta}$$

and

$$\beta = \beta(P,\nu) \;:=\; \int \frac{(\nu+q)\nu}{(\nu+\|x\|^2)^2}\,P(dx).$$

7. *M*-functionals of location and scatter

Now we return to the estimation of location and scatter as in Section 3.1. We restrict our attention to M-functionals derived from multivariate t-distributions with $\nu \geq 1$ degrees of freedom. That means, for an arbitrary distribution P on $\mathbb{R}^q$ we consider

$$L(\mu,\Sigma,P) \;:=\; \int \big[\rho\big((x-\mu)^\top\Sigma^{-1}(x-\mu)\big) - \rho(x^\top x)\big]\,P(dx) + \log\det(\Sigma)$$

as in (3.2), where

$$\rho(s) \;=\; \rho_{\nu,q}(s) := (\nu+q)\log(\nu+s).$$

The reason for the restriction to $\rho_{\nu,q}$ with $\nu \geq 1$ is a nice trick by Kent and Tyler (1991) to reduce the location-scatter problem in dimension q to the scatter-only problem in dimension $q+1$ with $\nu - 1$ in place of ν. As shown by Kent et al. (1994), the particular loss functions $\rho_{\nu,q}$ are the only ones for which this trick works.

For more details about and generalizations of multivariate t-distributions we refer to Lange et al. (1989) and the monograph by Kotz and Nadarajah (2004). An alternative approach to the location-scatter problem which is closely related to Tyler's (1987a) scatter functional is presented by Hettmansperger and Randles (2002).

7.1. Existence and uniqueness

The first question is under what conditions on P the functional $L(\cdot, \cdot, P)$ admits a unique minimizer $(\boldsymbol{\mu}(P), \boldsymbol{\Sigma}(P))$. To this end let

$$y = y(x) := \begin{bmatrix} x \\ 1 \end{bmatrix} \quad \text{and} \quad \Gamma := \begin{bmatrix} \Sigma + \mu\mu^\top & \mu \\ \mu^\top & 1 \end{bmatrix} = \begin{bmatrix} I_q & \mu \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} I_q & \mu \\ 0 & 1 \end{bmatrix}^\top$$

for $x \in \mathbb{R}^q$ and $(\mu, \Sigma) \in \mathbb{R}^q \times \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$. Then one can easily verify that

$$\det(\Gamma) = \det(\Sigma), \quad \Gamma^{-1} = \begin{bmatrix} I_q & -\mu \\ 0 & 1 \end{bmatrix}^\top \begin{bmatrix} \Sigma^{-1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} I_q & -\mu \\ 0 & 1 \end{bmatrix}$$

and

$$y^\top \Gamma^{-1} y = (x-\mu)^\top \Sigma^{-1} (x-\mu) + 1.$$

Consequently, with

$$\tilde{P} := \mathcal{L}(y(X)), \quad X \sim P,$$

and

$$\tilde{\rho}(s) := \rho(s-1) = \rho_{\nu-1,q+1}(s)$$

we may write

$$L(\mu, \Sigma, P) = \tilde{L}(\Gamma, \tilde{P}) := \int \left[\tilde{\rho}(y^\top \Gamma^{-1} y) - \tilde{\rho}(y^\top y)\right] \tilde{P}(dy) + \log\det(\Gamma).$$

If a matrix $\Gamma \in \mathbb{R}^{(q+1)\times(q+1)}_{\mathrm{sym},>0}$ minimizes $\tilde{L}(\cdot, \tilde{P})$, and if

$$\Gamma_{q+1,q+1} = 1,$$

then we may write

$$\Gamma = \begin{bmatrix} \boldsymbol{\Sigma}(P) + \boldsymbol{\mu}(P)\boldsymbol{\mu}(P)^\top & \boldsymbol{\mu}(P) \\ \boldsymbol{\mu}(P)^\top & 1 \end{bmatrix},$$

and $(\boldsymbol{\mu}(P), \boldsymbol{\Sigma}(P)) \in \mathbb{R}^q \times \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ solves the original minimization problem. It will turn out that the additional constraint $\Gamma_{q+1,q+1} = 1$ poses no problem here.

Concerning the minimization of $\tilde{L}(\cdot, \tilde{P})$ over $\mathbb{R}^{(q+1)\times(q+1)}_{\mathrm{sym},>0}$, one can deduce from Theorem 4.9 that the following condition on P plays a crucial role:

$$P(a+\mathbb{V}) \;<\; \frac{\dim(\mathbb{V})+\nu}{q+\nu} \quad \begin{array}{l}\text{for arbitrary } a\in\mathbb{R}^q \text{ and linear}\\ \text{subspaces } \mathbb{V}\subset\mathbb{R}^q \text{ with } 0\le\dim(\mathbb{V})<q.\end{array} \tag{7.1}$$

Here is the main result:

Theorem 7.1. *In case of $\nu = 1$, the functional $\tilde{L}(\cdot, \tilde{P})$ has a unique minimizer Γ with $\Gamma_{q+1,q+1} = 1$ if, and only if, (7.1) holds true. Moreover, if $\tilde{\Gamma}$ is some minimizer of $\tilde{L}(\cdot, \tilde{P})$, then $\Gamma = (\tilde{\Gamma}_{q+1,q+1})^{-1}\tilde{\Gamma}$.*

In case of $\nu > 1$, the functional $\tilde{L}(\cdot, \tilde{P})$ has a unique minimizer Γ if, and only if, (7.1) holds true. This minimizer satisfies automatically $\Gamma_{q+1,q+1} = 1$.

Consequently, Condition (7.1) is both necessary and sufficient for $L(\cdot,\cdot,P)$ to have a unique minimizer $(\boldsymbol{\mu}(P), \boldsymbol{\Sigma}(P))$. In that case, we have to minimize $\tilde{L}(\cdot, \tilde{P})$, which is equivalent to finding a solution $\Gamma \in \mathbb{R}^{(q+1)\times(q+1)}_{\mathrm{sym},>0}$ of the fixed point equation

$$\Gamma \;=\; \int \rho'(\|y\|^2 - 1)\, yy^\top \,\tilde{P}(dy) = \int \rho'(\|x\|^2) y(x)y(x)^\top \,P(dx).$$

If we write such a matrix Γ as

$$\Gamma \;=\; \begin{bmatrix} A & b \\ b^\top & c \end{bmatrix}$$

with $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$, $b \in \mathbb{R}^q$ and $c = \Gamma_{q+1,q+1} > 0$, then

$$\boldsymbol{\mu}(P) \;=\; c^{-1}b \quad \text{and} \quad \boldsymbol{\Sigma}(P) \;=\; c^{-1}A - \boldsymbol{\mu}(P)\boldsymbol{\mu}(P)^\top.$$

Moreover, $c = 1$ in case of $\nu > 1$.

7.2. Weak differentiability and linear expansions

The results for weak continuity and differentiability of scatter-only functionals imply analogous results for the location-scatter problem. Let $(P_n)_n$ be a sequence of probability distributions on $\mathbb{R}^q$ converging weakly to a distribution P such that $(\boldsymbol{\mu}(P), \boldsymbol{\Sigma}(P))$ is well-defined. Then for sufficiently large n, $(\boldsymbol{\mu}(P_n), \boldsymbol{\Sigma}(P_n))$ is well-defined, too, and

$$(\boldsymbol{\mu}(P_n), \boldsymbol{\Sigma}(P_n)) \;\to\; (\boldsymbol{\mu}(P), \boldsymbol{\Sigma}(P)).$$

(Again asymptotic statements are meant as $n \to \infty$.) This follows from Theorem 6.1, applied to $Q_{(n)} := \mathcal{L}\big(y(X)y(X)^\top\big)$, $X \sim P_{(n)}$. Theorem 6.4 yields the following expansion:

Theorem 7.2. *Let P be a probability distribution on $\mathbb{R}^q$ such that $\boldsymbol{\mu}(P) = 0$ and $\boldsymbol{\Sigma}(P) = I_q$. Then there exists a bounded and continuous function*

$$\tilde{Z} : \mathbb{R}^q \to \mathbb{R}^{(q+1)\times(q+1)}_{\mathrm{sym}}$$

depending only on P such that $\int \tilde{Z}\,dP = 0$ with the following property: Let $\widehat{P}_1$, $\widehat{P}_2$, $\widehat{P}_3$, ... be random distributions on $\mathbb{R}^q$ such that for any bounded and continuous function $f : \mathbb{R}^q \to \mathbb{R}$,

$$\int f\,d\widehat{P}_n \;\to_p\; \int f\,dP.$$

Then $\big(\boldsymbol{\mu}(\widehat{P}_n), \boldsymbol{\Sigma}(\widehat{P}_n)\big)$ is well-defined with asymptotic probability one, and

$$\begin{bmatrix} \boldsymbol{\Sigma}(\widehat{P}_n) - I_q & \boldsymbol{\mu}(\widehat{P}_n) \\ \boldsymbol{\mu}(\widehat{P}_n)^\top & 0 \end{bmatrix} \;=\; \int \big(\tilde{Z} - \tilde{Z}_{q+1,q+1} I_{q+1}\big)\,d\widehat{P}_n + o_p\Big(\Big\|\int \tilde{Z}\,d\widehat{P}_n\Big\|\Big).$$

The precise definition of $\tilde{Z}$ is

$$\tilde{Z}(x) \;:=\; \tilde{H}(P)^{-1}\big(\rho'(\|x\|^2) y(x) y(x)^\top - I_{q+1}\big),$$

where $\tilde{H}(P) : \tilde{\mathbb{M}} \to \tilde{\mathbb{M}}$ is the linear operator given by

$$\tilde{H}(P)M \;:=\; M + \int \rho''(\|x\|^2)\, y(x)^\top M y(x)\, y(x) y(x)^\top\, P(dx)$$

for matrices M in

$$\tilde{\mathbb{M}} \;:=\; \begin{cases} \big\{M \in \mathbb{R}^{(q+1)\times(q+1)}_{\mathrm{sym}} : \mathrm{tr}(M) = 0\big\} & \text{if } \nu = 1, \\ \mathbb{R}^{(q+1)\times(q+1)}_{\mathrm{sym}} & \text{if } \nu > 1. \end{cases}$$

Moreover, in case of $\nu > 1$,

$$\tilde{Z}_{q+1,q+1} \;\equiv\; 0.$$

Remark 7.3 (Empirical distributions)**.** Let $P_1, P_2, P_3, \ldots$ and P be distributions on $\mathbb{R}^q$ such that $P_n \to_w P$ and $\boldsymbol{\mu}(P) = 0 = \boldsymbol{\mu}(P_n)$ and $\boldsymbol{\Sigma}(P) = I_q = \boldsymbol{\Sigma}(P_n)$ for all n. Further let $\widehat{P}_n$ be the empirical distribution of independent random vectors $X_{n1}, X_{n2}, \ldots, X_{nn}$ with distribution P_n. As in the proof of Theorem 6.10 one can show that these random distributions $\widehat{P}_n$ satisfy the assumptions of Theorem 7.2. This implies that $(\boldsymbol{\mu}(\widehat{P}_n), \boldsymbol{\Sigma}(\widehat{P}_n))$ is well-defined with asymptotic probability one, and

$$\sqrt{n}\begin{bmatrix} \boldsymbol{\Sigma}(\widehat{P}_n) - I_q & \boldsymbol{\mu}(\widehat{P}_n) \\ \boldsymbol{\mu}(\widehat{P}_n)^\top & 0 \end{bmatrix} \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^n \big(\tilde{Z}(X_{ni}) - \tilde{Z}(X_{ni})_{q+1,q+1} I_q\big) + o_p(1)$$

with $\tilde{Z} : \mathbb{R}^q \to \mathbb{R}^{(q+1)\times(q+1)}_{\mathrm{sym}}$ as in Theorem 7.2. In particular, $\mathbb{E}\tilde{Z}(X_{n1}) = 0$ for all n, and the random matrix in the previous display converges in distribution to a random matrix with a centered Gaussian distribution on $\mathbb{R}^{(q+1)\times(q+1)}_{\mathrm{sym}}$.

Remark 7.4 (Symmetry)**.** Suppose that P is symmetric in the sense that $\mathcal{L}(-X) = \mathcal{L}(X)$ for $X \sim P$. Then the function $\tilde{Z}$ in Theorem 7.2 may be written as

$$\tilde{Z}(x) \;=\; \begin{bmatrix} Z(xx^\top) & 0 \\ 0 & z(\|x\|^2) \end{bmatrix} + \rho'(\|x\|^2) \begin{bmatrix} 0 & Bx \\ x^\top B & 0 \end{bmatrix}$$

with bounded and continuous functions $Z : \mathbb{R}^{q\times q}_{\mathrm{sym},\geq 0} \to \mathbb{R}^{q\times q}_{\mathrm{sym}}$, $z : [0,\infty) \to \mathbb{R}$ and a nonsingular matrix $B \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$. In particular, the random variables $\sqrt{n}(\boldsymbol{\Sigma}(\widehat{P}_n) - I_q)$ and $\sqrt{n}\boldsymbol{\mu}(\widehat{P}_n)$ in Remark 7.3 are asymptotically independent.

Remark 7.5 (Spherical symmetry)**.** Suppose that P is spherically symmetric around 0. Let $\beta = \beta(P,\nu)$, $A_0(\cdot)$ and $a(\cdot)$ be defined as in Remark 6.14. Then the function $\tilde{Z} - \tilde{Z}_{q+1,q+1}I_{q+1}$ in Theorem 7.2 may be written as follows:

$$\tilde{Z}(x) - \tilde{Z}(x)_{q+1,q+1}I_{q+1} \;=\; (\nu + \|x\|^2)^{-1}\begin{bmatrix} c_0A_0(x) + c_1a(x)I_q & c_2x \\ c_2x^\top & 0 \end{bmatrix}$$

where

$$c_0 \;:=\; \frac{(q+\nu)(q+2)}{q+2(1-\beta)\nu/q}, \quad c_1 \;:=\; \frac{q}{1-\beta} \quad \text{and} \quad c_2 \;:=\; \frac{q}{q-2(1-\beta)}.$$

Comparing this with Remark 6.14, we see that the estimator $\boldsymbol{\Sigma}(\widehat{P}_n)$ has the same asymptotic behaviour as the corresponding estimator in the scatter-only problem.

8. Auxiliary results and proofs

8.1. Proofs for Section 2

Proof of Lemma 2.4. Note that $(X_{\pi(i)})_{i=1}^q = BX$ with the permutation matrix $B = (1_{[\pi(i)=j]})_{i,j=1}^q$. Thus our assumption on X in part (i) and linear equivariance of $\boldsymbol{\Sigma}(\cdot)$ imply that

$$\boldsymbol{\Sigma}(P) \;=\; B\boldsymbol{\Sigma}(P)B^\top \;=\; \big(\boldsymbol{\Sigma}(P)_{\pi(i),\pi(j)}\big)_{i,j=1}^q$$

for any permutation π of $\{1,2,\ldots,q\}$ such that $\pi(i) = i$ whenever $i \notin J$. Let $j_1 := \min(J)$ and $j_2 := \max(J)$. For arbitrary indices $j \neq k$ in J, choose π such that $\pi(j_1) = j$ and $\pi(j_2) = k$. Then we realize that $\boldsymbol{\Sigma}(P)_{j,j} = a(P) := \boldsymbol{\Sigma}(P)_{j_1,j_1}$ and $\boldsymbol{\Sigma}(P)_{j,k} = b(P) := \boldsymbol{\Sigma}(P)_{j_1,j_2}$. This proves part (i).

To verify part (ii) we write $(s_iX_i)_{i=1}^q = BX$ with $B := \mathrm{diag}(s)$. Then

$$\boldsymbol{\Sigma}(P) \;=\; B\boldsymbol{\Sigma}(P)B^\top \;=\; \big(s_is_j\boldsymbol{\Sigma}(P)_{i,j}\big)_{i,j=1}^q.$$

Consequently, $\boldsymbol{\Sigma}(P)_{ij} = 0$ whenever $s_is_j = -1$, i.e. $s_i \neq s_j$.

As for part (iii), suppose first that P is spherically symmetric. This implies that $X \sim P$ satisfies the assumptions of part (i) with the full index set $J = \{1,2,\ldots,q\}$ and of part (ii) for any sign vector $s \in \{-1,1\}^q$. Hence $\boldsymbol{\Sigma}(P) = c(P)I_q$ for some $c(P) \geq 0$. Now suppose that P is elliptically symmetric with center 0 and scatter matrix $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$. Then the distribution P' of $X' := \Sigma^{-1/2}X$ is spherically symmetric, and $P = {P'}^B$ with $B := \Sigma^{1/2}$. Thus $\boldsymbol{\Sigma}(P) = B\boldsymbol{\Sigma}(P')B^\top = c(P')\Sigma$. □

Proof of Lemma 2.5. Under the assumption of part (i),

$$\boldsymbol{\mu}(P) \;=\; \mathrm{diag}(s)\boldsymbol{\mu}(P) \;=\; \big(s_i\boldsymbol{\mu}(P)_i\big)_{i=1}^q.$$

Consequently, $\boldsymbol{\mu}(P)_i = 0$ whenever $s_i = -1$.

If P is elliptically symmetric with center μ and scatter matrix Σ, then the distribution P' of $X' := \Sigma^{-1/2}(X-\mu)$ is spherically symmetric, and $P = P'^{\mu,B}$ with $B := \Sigma^{1/2}$. But X' satisfies the assumptions of part (i) for any sign vector $s \in \{-1,1\}^q$. Hence $\boldsymbol{\mu}(P') = 0$, and $\boldsymbol{\mu}(P) = \mu + B\boldsymbol{\mu}(P') = \mu$. Moreover, $\boldsymbol{\Sigma}(P) = B\boldsymbol{\Sigma}(P')B^\top = c(P')\Sigma$, according to Lemma 2.4, applied to P'. □

8.2. *Proofs for Section 4*

Proof of Lemma 4.7. Let $M = \sum_{i=1}^q \lambda_i(M) u_i u_i^\top$ with eigenvalues $\lambda_i(M) \geq 0$ and an orthonormal basis $u_1, u_2, \ldots, u_q$ of $\mathbb{R}^q$. Then $\operatorname{tr}(M) = \sum_{i=1}^q \lambda_i(M)$ and

$$\operatorname{tr}(AM) \;=\; \sum_{i=1}^q \lambda_i(M) u_i^\top A u_i \quad \begin{cases} \leq\; \lambda_{\max}(A) \sum_{i=1}^q \lambda_i(M) \;=\; \lambda_{\max}(A)\operatorname{tr}(M), \\ \geq\; \lambda_{\min}(A) \sum_{i=1}^q \lambda_i(M) \;=\; \lambda_{\min}(A)\operatorname{tr}(M). \end{cases}$$ □

Proof of Lemma 4.8. For fixed $s > 0$ and $x \in \mathbb{R}$ define $f(x) := \rho(e^x s)$. Then $f'(x) = \rho'(e^x s)e^x s = \psi(e^x s)$. Consequently by the mean value theorem,

$$\rho(t) - \rho(s) \;=\; f(\log(t/s)) - f(0) \;=\; f'(\xi)\log(t/s) \;=\; \psi(e^\xi s)\log(t/s)$$

with some number ξ between 0 and $\log(t/s)$. Since ψ is non-decreasing on $(0,\infty)$, either $\log(t/s) > 0$ and $\psi(s) \leq \psi(e^\xi s) \leq \psi(t)$, or $\log(t/s) < 0$ and $\psi(t) \leq \psi(e^\xi s) \leq \psi(s)$. In both cases, $\psi(s)\log(t/s) \leq \rho(t) - \rho(s) \leq \psi(t)\log(t/s)$.

Note also that

$$\rho(t) - \rho(s) \;=\; \rho'(\xi)(t-s)$$

for some ξ between a and b. Hence if ρ' is non-increasing, the asserted inequalities follow from the fact that either $t - s \geq 0$ and $\rho'(t) \leq \rho'(\xi) \leq \rho'(s)$, or $t - s < 0$ and $\rho'(s) \leq \rho'(\xi) \leq \rho'(t)$. □

Proof of Lemma 4.10. It follows from (4.8) that

$$\mathbb{P}\big(X_1, X_2, \ldots, X_k \text{ are linearly independent}\big) \;=\; 1 \quad \text{for } k = 1, 2, \ldots, q.$$

Indeed, $\mathbb{P}(X_1 \neq 0) = 1$, and for $2 \leq k \leq q$,

$$\mathbb{P}\big(X_k \notin \operatorname{span}(X_1, \ldots, X_{k-1}) \,\big|\, X_1, \ldots, X_{k-1}\big) \;=\; 1.$$

This implies that with probability one,

$$\widehat{Q}^1(\mathbb{M}(\mathbb{V})) = \widehat{P}(\mathbb{V}) \;\leq\; \frac{\dim(\mathbb{V})}{n} \quad \text{for all } \mathbb{V} \in \mathcal{V}_q \text{ with } \dim(\mathbb{V}) < q.$$

Consequently, according to Theorem 4.9, $\boldsymbol{\Sigma}_\rho(\widehat{Q}^1)$ is well-defined with probability one, provided that

$$\frac{d}{n} \;<\; \begin{cases} \dfrac{d}{q} & \text{for } 1 \leq d < q, \quad \text{in Case 0,} \\ \dfrac{\psi(\infty) - q + d}{\psi(\infty)} & \text{for } 0 \leq d < q, \quad \text{in Case 1.} \end{cases}$$

But this can be shown to be equivalent to $n \geq q+1$ in Case 0 and $n \geq q$ in Case 1. □

To understand setting (4.2) thoroughly, the following two results about linear subspaces of $\mathbb{R}^q$ and sample covariance matrices are useful:

Lemma 8.1. *For arbitrary integers $k \geq 1$ and points $x_1, x_2, \ldots, x_k \in \mathbb{R}^q$ with sample mean $\bar{x} = k^{-1} \sum_{i=1}^k x_i$,*

$$\begin{aligned} \mathbb{W}(x_1, x_2, \ldots, x_k) &:= \operatorname{span}(x_i - x_j \,:\, i, j = 1, 2, \ldots, k) \\ &= \operatorname{span}(x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_k - \bar{x}) \\ &= \operatorname{span}(x_1 - x_a, x_2 - x_a, \ldots, x_k - x_a) \end{aligned}$$

for any $a \in \{1, 2, \ldots, k\}$. Moreover, in case of $k \geq 2$,

$$S(x_1, x_2, \ldots, x_k)\, \mathbb{R}^q \;=\; \mathbb{W}(x_1, x_2, \ldots, x_k).$$

Corollary 8.2. *Let $x_1, x_2, \ldots, x_k$ and $y_1, y_2, \ldots, y_\ell$ be arbitrary point in $\mathbb{R}^q$. Suppose that both $\mathbb{W}(x_1, x_2, \ldots, x_k)$ and $\mathbb{W}(y_1, y_2, \ldots, y_\ell)$ are contained in a given space $\mathbb{V} \in \mathcal{V}_q$. If $\{x_1, x_2, \ldots, x_k\}$ and $\{y_1, y_2, \ldots, y_\ell\}$ have at least one point in common, then*

$$\mathbb{W}(x_1, x_2, \ldots, x_m,\, y_1, y_2, \ldots, y_\ell) \;\subset\; \mathbb{V}.$$

Proof of Lemma 8.1. For arbitrary indices $a, j \in \{1, 2, \ldots, k\}$ we may write $x_j - \bar{x} = (x_j - x_a) - k^{-1} \sum_{i=1}^k (x_i - x_a)$, so

$$\begin{aligned} &\operatorname{span}(x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_k - \bar{x}) \\ &\quad\subset \operatorname{span}(x_1 - x_a, x_2 - x_a, \ldots, x_k - x_a) \\ &\quad\subset \operatorname{span}(x_i - x_j \,:\, i, j = 1, 2, \ldots, k) \\ &\quad= \operatorname{span}\big((x_i - \bar{x}) - (x_j - \bar{x}) \,:\, i, j = 1, 2, \ldots, k\big) \\ &\quad\subset \operatorname{span}(x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_k - \bar{x}). \end{aligned}$$

Hence the preceding three inclusions are equalities.

Now suppose that $k \geq 2$. Since $S := S(x_1, x_2, \ldots, x_k)$ is positive semidefinite, it follows from its spectral representation that a vector $w \in \mathbb{R}^q$ is perpendicular to the column space $S\,\mathbb{R}^q$ if, and only if,

$$0 \;=\; w^\top S w \;=\; (k-1)^{-1} \sum_{i=1}^k (w^\top (x_i - \bar{x}))^2,$$

i.e. w is perpendicular to $\operatorname{span}(x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_k - \bar{x}) = \mathbb{W}(x_1, x_2, \ldots, x_k)$. Hence the column space of S is equal to $\mathbb{W}(x_1, x_2, \ldots, x_k)$. □

Proof of Lemma 4.11. For any nonvoid index set $M \subset \{1, 2, \ldots, n\}$ define $\mathbb{W}(M) := \mathbb{W}(X_i : i \in M)$; in particular, $\mathbb{W}(\{i\}) = \{0\}$. Then it follows from Lemma 8.1 that for any $\mathbb{V} \in \mathcal{V}_q$,

$$\widehat{Q}^k(\mathbb{M}(\mathbb{V})) \;=\; \binom{n}{k}^{-1} \sum_{J \in \mathcal{J}_k} 1_{[\mathbb{W}(J) \subset \mathbb{V}]},$$

where $\mathcal{J}_k$ stands for the set of all subsets of $\{1,2,\ldots,n\}$ with k elements. Moreover, Corollary 8.2 implies that for two nonvoid index sets M, M',

$$\mathbb{W}(M \cup M') \ \subset \mathbb{V} \quad \text{if} \ \ \mathbb{W}(M) \subset \mathbb{V}, \mathbb{W}(M') \subset \mathbb{V} \text{ and } M \cap M' \neq \emptyset.$$

Consequently, if we partition $\{1,2,\ldots,n\}$ into pairwise disjoint and maximal subsets $M_1, M_2, \ldots, M_L$ such that $\mathbb{W}(M_\ell) \subset \mathbb{V}$ for $\ell = 1,2,\ldots,L$, then

$$\widehat{Q}^k(\mathbb{M}(\mathbb{V})) \ = \ \binom{n}{k}^{-1} \sum_{\ell=1}^{L} \binom{\#M_\ell}{k}$$

with the usual convention that $\binom{a}{k} := 0$ for integers $0 \le a < k$.

For any fixed index set M with $1 \le \#M \le q$ and an additional index $j \notin M$, it follows from (4.10) and Lemma 8.1 that

$$\begin{aligned} \mathbb{P}\big(\mathbb{W}(M \cup \{j\}) \neq \mathbb{W}(M) \,\big|\, (X_i)_{i \neq j}\big) \ &= \ \mathbb{P}\big(X_j - X_a \notin \mathbb{W}(M) \,\big|\, (X_i)_{i \neq j}\big) \\ &= \ 1, \end{aligned}$$

where a is any index in M. This implies that with probability one, for any given partition $M_1, M_2, \ldots, M_L$ of $\{1,2,\ldots,n\}$ into nonvoid subsets M_ℓ,

$$\dim\Big(\bigcup_{\ell=1}^{L} \mathbb{W}(M_\ell)\Big) \ = \ \min\Big(\sum_{\ell=1}^{L} (\#M_\ell - 1), q\Big).$$

In particular, for any $\mathbb{V} \in \mathcal{V}_q$ with $d := \dim(\mathbb{V}) < q$, the value of $\widehat{Q}^k(\mathbb{M}(\mathbb{V}))$ is no larger than the maximum of

$$\binom{n}{k}^{-1} \sum_{\ell=1}^{L} \binom{m_\ell + 1}{k} \tag{8.1}$$

over all integers $L \ge 1$ and $m_1, m_2, \ldots, m_L \ge k-1$ such that $\sum_{\ell=1} m_\ell \le d$. It will be shown later that this maximum equals

$$\binom{n}{k}^{-1} \binom{d+1}{k}.$$

Since $(\psi(\infty) - q + d)/\psi(\infty) = 1 - (q-d)/\psi(\infty) > 1 - (q-d)/q = d/q$ in Case 1, we conclude that $\mathbf{\Sigma}_\rho(\widehat{Q}^k)$ is well-defined almost surely, provided that

$$\binom{n}{k}^{-1} \binom{d+1}{k} \ < \ \frac{d}{q} \quad \text{for } k-1 \le d < q.$$

Since $\binom{d+1}{k}/d$ is increasing in $d \ge k-1$, this condition is equivalent to

$$\binom{n}{k}^{-1} \binom{q}{k} \ < \ \frac{q-1}{q}.$$

But this holds in case of $n \ge q+1$, since the left hand side equals

$$\binom{n}{k}^{-1} \binom{q}{k} \ \le \ \binom{q+1}{k}^{-1} \binom{q}{k} \ = \ \frac{q-k+1}{q+1} \ \le \ \frac{q-1}{q+1} \ < \ \frac{q-1}{q}.$$

It remains to be shown that the sum $\sum_{\ell=1}^{L}\binom{m_\ell+1}{k}$ in (8.1) is not larger than $\binom{d+1}{k}$. For this purpose, let $N_1, N_2, \ldots, N_L$ be disjoint subsets of $\{1,2,\ldots,d\}$ with $\#N_\ell = m_\ell$, and let $M_\ell := N_\ell \cup \{d+1\}$. Then for $\ell, \ell' \in \{1,2,\ldots,L\}$ with $\ell \neq \ell'$, a subset of M_ℓ with k elements is different from any subset of $M_{\ell'}$ with k elements. Consequently,

$$\begin{aligned}\sum_{\ell=1}^{L}\binom{m_\ell+1}{k} &= \sum_{\ell=1}^{L}\#\big\{\text{subsets of } M_\ell \text{ with } k \text{ elements}\big\}\\ &\le \#\big\{\text{subsets of } \{1,2,\ldots,d+1\} \text{ with } k \text{ elements}\big\}\\ &= \binom{d+1}{k}.\end{aligned}$$

□

Proof of Theorem 4.9. The first part, i.e. the equivalence of the fixed-point equation $\Psi_\rho(\Sigma, Q) = \Sigma$ and Σ being a minimizer of $L_\rho(\cdot, Q)$, follows from Propositions 5.2 and 5.4: Recall that with $B := \Sigma^{1/2}$ we may write

$$\begin{aligned}L_\rho(\Sigma^{1/2}\exp(A)\Sigma^{1/2}, Q) - L_\rho(\Sigma, Q) &= L_\rho(\exp(A), Q_B)\\ &= \big\langle A, G_\rho(Q_B)\big\rangle + o(\|A\|)\end{aligned}$$

as $\mathbb{R}^{q\times q}_{\mathrm{sym}} \ni A \to 0$, and

$$G_\rho(Q_B) = B^{-1}(\Sigma - \Psi_\rho(\Sigma, Q))B^{-1}.$$

If Σ minimizes $L_\rho(\cdot, Q)$, then $G_\rho(Q_B) = 0$, which is equivalent to $\Psi_\rho(\Sigma, Q) = \Sigma$. On the other hand, if Σ is not a minimizer of $L_\rho(\cdot, Q)$, then there exists a matrix $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$ such that $L_\rho(\exp(A), Q_B) < 0$. But convexity of $\mathbb{R} \ni t \mapsto h(t) := L_\rho(\exp(tA), Q_B)$ implies that

$$0 > L_\rho(\exp(A), Q_B) = h(1) - h(0) \ge h'(0) = \langle A, G_\rho(Q_B)\rangle,$$

i.e. $G_\rho(Q_B) \neq 0$ and thus $\Psi_\rho(\Sigma, Q) \neq \Sigma$.

In Case 1, suppose that Condition 1 holds true. According to Proposition 5.5, $L(\cdot, Q)$ is a continuous function on $\mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ which is coercive in that $L_\rho(\Sigma, Q) \to \infty$ as $\|\log(\Sigma)\| \to \infty$. Consequently there exists a minimizer Σ_o of $L_\rho(\cdot, Q)$. But Condition 1 and Proposition 5.4 imply that $L_\rho(\Sigma_o^{1/2}\exp(tA)\Sigma_o^{1/2}, Q)$ is strictly convex for any $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}} \setminus \{0\}$. Consequently, Σ_o is the unique minimizer of $L_\rho(\cdot, Q)$.

Still in Case 1, suppose that $\Sigma_o \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ is a unique minimizer of $L_\rho(\cdot, Q)$. Then $L_\rho(\Sigma_o^{1/2}\exp(A)\Sigma_o^{1/2}, Q)$ is a coercive function of $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$: For if $\|A\| \ge 1$ and $A' := \|A\|^{-1}A$, then by Proposition 5.4,

$$\begin{aligned}&L_\rho(\Sigma_o^{1/2}\exp(A)\Sigma_o^{1/2}, Q) - L_\rho(\Sigma_o, Q)\\ &\quad= L_\rho(\Sigma_o^{1/2}\exp(\|A\|A')\Sigma_o^{1/2}, Q) - L_\rho(\Sigma_o, Q)\\ &\quad\ge \|A\|\big(L_\rho(\Sigma_o^{1/2}\exp(A')\Sigma_o^{1/2}, Q) - L_\rho(\Sigma_o, Q)\big)\\ &\quad\ge \|A\| \min_{A''\in\mathbb{R}^{q\times q}_{\mathrm{sym}}\,:\,\|A''\|=1}\big(L_\rho(\Sigma_o^{1/2}\exp(A'')\Sigma_o^{1/2}, Q) - L_\rho(\Sigma_o, Q)\big),\end{aligned}$$

and the minimum on the right hand side is strictly positive by uniqueness of the minimizer Σ_o. But coercivity of $L_\rho(\Sigma_o^{1/2}\exp(\cdot)\Sigma_o^{1/2}, Q)$ is equivalent to Condition 1, according to Proposition 5.5.

In Case 0 one can argue in the same way, this time with $\{\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0} : \det(\Sigma) = 1\}$ and $\{A \in \mathbb{R}^{q\times q}_{\mathrm{sym}} : \mathrm{tr}(A) = 0\}$ in place of $\mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ and $\mathbb{R}^{q\times q}_{\mathrm{sym}}$, respectively. □

Proof of Lemma 4.13. Writing $\Psi(\tilde{Q}) = \Psi_\rho(I, \tilde{Q})$ for arbitrary distributions $\tilde{Q}$ and $B := \Sigma^{1/2}$, note first that

$$\begin{aligned} L_\rho(\Psi_\rho(\Sigma, Q), Q) - L_\rho(\Sigma, Q) &= L_\rho(B\Psi(Q_B)B^\top, Q) - L_\rho(BB^\top, Q) \\ &= L_\rho(\Psi(Q_B), Q_B). \end{aligned}$$

Hence it suffices to show that

$$L_\rho(\Psi(Q_B), Q_B) < 0$$

unless $\Psi(Q_B) = I_q$. It follows from the second part of Lemma 4.8 that for $\Gamma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$,

$$\begin{aligned} L_\rho(\Gamma, Q_B) &= \int \big[\rho(\mathrm{tr}(\Gamma^{-1}M)) - \rho(\mathrm{tr}(M))\big]\, Q_B(dM) + \log\det(\Gamma) \\ &\le \int \rho'(\mathrm{tr}(M))\big[\mathrm{tr}(\Gamma^{-1}M) - \mathrm{tr}(M)\big]\, Q_B(dM) + \log\det(\Gamma) \\ &= \mathrm{tr}\big((\Gamma^{-1} - I_q)\Psi(Q_B)\big) + \log\det(\Gamma). \end{aligned}$$

Hence

$$\begin{aligned} L_\rho(\Psi(Q_B), Q_B) &\le \mathrm{tr}(I_q - \Psi(Q_B)) + \log\det\Psi(Q_B) \\ &= \sum_{i=1}^q \big[1 - \lambda_i(\Psi(Q_B)) + \log\lambda_i(\Psi(Q_B))\big]. \end{aligned}$$

Since $1 - x + \log x < 0$ for $0 < x \ne 1$, the latter sum is strictly negative unless $\lambda_i(\Psi(Q_B)) = 1$ for $1 \le i \le q$, which is equivalent to $\Psi(Q_B) = I_q$, i.e. $\Psi_\rho(\Sigma, Q) = \Sigma$. □

Proof of Lemma 4.12. Under the stated conditions on the distribution Q, the function $L_\rho(\cdot, Q)$ has a minimizer Σ_o, that means, $\Psi_\rho(\Sigma_o, Q) = \Sigma_o$. Note that

$$\Sigma_o^{-1/2}\Sigma_k\Sigma_o^{-1/2} = \Sigma_o^{-1/2}\Psi_\rho(\Sigma_{k-1}, Q)\Sigma_o^{-1/2} = \Psi_\rho(\Sigma_o^{-1/2}\Sigma_{k-1}\Sigma_o^{-1/2}, Q_{\Sigma_o^{1/2}}).$$

Hence we may assume w.l.o.g. that $\Sigma_o = I_q$ and $\Psi_\rho(I_q, Q) = I_q$. Again we write $\Psi(\cdot)$ instead of $\Psi_\rho(\cdot, Q)$.

The equation $\Psi(I_q) = I_q$ implies that the mapping Ψ has the following properties, as shown below: For any $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$,

$$\lambda_{\min}(\Psi(\Sigma)) \ge a := \begin{cases} \lambda_{\min}(\Sigma) & \text{in Case 0,} \\ \min\{\lambda_{\min}(\Sigma), 1\} & \text{in Case 1,} \end{cases}$$

$$\lambda_{\max}(\Psi(\Sigma)) \;\le\; b := \begin{cases} \lambda_{\max}(\Sigma) & \text{in Case 0,} \\ \max\{\lambda_{\max}(\Sigma), 1\} & \text{in Case 1.} \end{cases}$$

This follows from Lemma 4.7 and various properties of ρ: For any $M \in \mathbb{R}^{q\times q}_{\mathrm{sym},\ge 0}$,

$$\lambda_{\max}(\Sigma)^{-1}\operatorname{tr}(M) \;\le\; \operatorname{tr}(\Sigma^{-1}M) \;\le\; \lambda_{\min}(\Sigma)^{-1}\operatorname{tr}(M).$$

Hence for any unit vector $v \in \mathbb{R}^q$,

$$\begin{aligned} v^\top\Psi(\Sigma)v \;&=\; \int \rho'(\operatorname{tr}(\Sigma^{-1}M))\, v^\top M v\, Q(dM) \\ &\begin{cases} \ge\; a\displaystyle\int \rho'(\operatorname{tr}(M))\, v^\top M v\, Q(dM) \;=\; a\, v^\top\Psi(I_q)v \;=\; a, \\ \le\; b\displaystyle\int \rho'(\operatorname{tr}(M))\, v^\top M v\, Q(dM) \;=\; b\, v^\top\Psi(I_q)v \;=\; b, \end{cases} \end{aligned}$$

because for $M \neq 0$,

$$\rho'(\operatorname{tr}(\Sigma^{-1}M)) \begin{cases} \ge \rho'(\operatorname{tr}(M)/a) = \dfrac{\psi(\operatorname{tr}(M)/a)}{\operatorname{tr}(M)/a} \ge \dfrac{a\psi(\operatorname{tr}(M))}{\operatorname{tr}(M)} = a\rho'(\operatorname{tr}(M)), \\ \le \rho'(\operatorname{tr}(M)/b) = \dfrac{\psi(\operatorname{tr}(M)/b)}{\operatorname{tr}(M)/b} \le \dfrac{b\psi(\operatorname{tr}(M))}{\operatorname{tr}(M)} = b\rho'(\operatorname{tr}(M)), \end{cases}$$

due to ρ' being non-increasing and ψ being constant in Case 0 and increasing on $(0,\infty)$ in Case 1.

Now we define

$$[a_k, b_k] \;:=\; \begin{cases} \big[\lambda_{\min}(\Sigma_k), \lambda_{\max}(\Sigma_k)\big] & \text{in Case 0,} \\ \big[\min\{\lambda_{\min}(\Sigma_k), 1\}, \max\{\lambda_{\max}(\Sigma_k), 1\}\big] & \text{in Case 1.} \end{cases}$$

Then $(a_k)_k$ and $(b_k)_k$ are non-decreasing and non-increasing, respectively, with corresponding limits $a_* \le b_*$. In Case 0 we have to show that $a_* = b_*$, because then $\Sigma_k \to a_* I_q$. In Case 1 we have to show that $a_* = b_* = 1$, because then $\Sigma_k \to I_q$. To this end, note that the set $\big\{\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym}} : \lambda(\Sigma) \in [a_0, b_0]^q\big\}$ is compact. Hence there exist indices $k(1) < k(2) < k(3) < \cdots$ such that $\Sigma_{k(\ell)} \to \Sigma_*$ as $\ell \to \infty$, where $\lambda(\Sigma_*) \in [a_0, b_0]^q$. Lemma 4.13 entails that the sequence $\big(L_\rho(\Sigma_k, Q)\big)_{k\ge 0}$ is non-increasing. Consequently, since $L_\rho(\cdot, Q)$ and $\Psi(\cdot)$ are continuous,

$$\begin{aligned} L_\rho(\Sigma_*, Q) \;&=\; \lim_{\ell\to\infty} L_\rho(\Sigma_{k(\ell)}, Q) \\ &=\; \lim_{\ell\to\infty} L_\rho(\Sigma_{k(\ell)+1}, Q) \;=\; \lim_{\ell\to\infty} L_\rho(\Psi(\Sigma_{k(\ell)}), Q) \;=\; L_\rho(\Psi(\Sigma_*), Q). \end{aligned}$$

Hence Lemma 4.13 implies that $\Psi(\Sigma_*) = \Psi_\rho(\Sigma_*, Q) = \Sigma_*$. Thus Σ_* is a minimizer of $L_\rho(\cdot, Q)$. In Case 0 this implies that Σ_* is a positive multiple of I_q, whence $a_* = \lambda_{\min}(\Sigma_*) = \lambda_{\max}(\Sigma_*) = b_*$. In Case 1 this implies that $\Sigma_* = I_q$, whence $a_* = \min\{\lambda_{\min}(\Sigma_*), 1\} = 1$ and $b_* = \max\{\lambda_{\max}(\Sigma_*), 1\} = 1$. □

8.3. Proofs for Section 5

Proof of Lemma 5.1. By definition,

$$\exp(A+\Delta) \;=\; \sum_{\ell=0}^{\infty} \frac{(A+\Delta)^\ell}{\ell!},$$

and for $\ell \geq 1$, the expansion of $(A+\Delta)^\ell$ is the sum of A^ℓ and all matrices of the form $A^{s_0}\Delta A^{s_1}\cdots\Delta A^{s_k}$ with $k \in \{1,\ldots,\ell\}$ times the factor Δ and exponents $s_0,\ldots,s_k \geq 0$ such that $s_+ := \sum_{j=0}^k s_j$ equals $\ell - k$. Consequently,

$$\exp(A+\Delta) \;=\; \exp(A) + \sum_{k=1}^{\infty} T_k(A,\Delta)$$

with

$$T_k(A,\Delta) \;:=\; \sum_{s_0,\ldots,s_k \geq 0} \frac{A^{s_0}\Delta A^{s_1}\cdots\Delta A^{s_k}}{(s_+ + k)!}.$$

Note that for given $\ell \geq k$ there are $\binom{\ell}{k}$ tupels $(s_0,\ldots,s_k)$ of integers $s_j \geq 0$ with $s_+ = \ell - k$. Thus

$$\|T_k(A,\Delta)\| \;\leq\; \sum_{\ell=k}^{\infty} \binom{\ell}{k} \frac{\|A\|^{\ell-k}\|\Delta\|^k}{\ell!} \;=\; e^{\|A\|}\frac{\|\Delta\|^k}{k!}.$$

In particular,

$$\exp(A+\Delta) \;=\; \exp(A) + R_0(A,\Delta) \;=\; \exp(A) + T_1(A,\Delta) + R_1(A,\Delta)$$

with

$$\|R_m(A,\Delta)\| \;\leq\; \sum_{k=m+1}^{\infty} e^{\|A\|}\frac{\|\Delta\|^k}{k!} \;\leq\; e^{\|A\|+\|\Delta\|}\frac{\|\Delta\|^{m+1}}{(m+1)!}$$

for $m = 0, 1$.

It remains to derive alternative expressions for $T_k(A,\Delta)$. First of all, it follows from a well-known identity for the beta function that

$$\begin{aligned}
T_1(A,\Delta) \;&=\; \sum_{s_0,s_1\geq 0} \frac{A^{s_0}\Delta A^{s_1}}{(s_0+s_1+1)!} \\
&=\; \sum_{s_0,s_1\geq 0} \frac{s_0!s_1!}{(s_0+s_1+1)!}\frac{A^{s_0}}{s_0!}\Delta\frac{A^{s_1}}{s_1!} \\
&=\; \sum_{s_0,s_1\geq 0} \int_0^1 (1-u)^{s_0}u^{s_1}\,du\;\frac{A^{s_0}}{s_0!}\Delta\frac{A^{s_1}}{s_1!} \\
&=\; \int_0^1 \sum_{s_0,s_1\geq 0} \frac{((1-u)A)^{s_0}}{s_0!}\Delta\frac{(uA)^{s_1}}{s_1!}\;du \\
&=\; \int_0^1 \exp((1-u)A)\Delta\exp(uA)\,du.
\end{aligned}$$

For general $k \geq 1$ we utilize a special construction of the random tupel $(U_{kj})_{j=0}^k$ which is well-known from uniform order statistics: If $E_0, E_1, E_2, \ldots$ are independent standard exponential random variables, then the random variable $(U_{kj})_{j=0}^k := (E_j/F)_{j=0}^k$ with $F := \sum_{j=0}^k E_j$ has the desired distribution. Moreover, $(U_{kj})_{j=0}^k$ and F are stochastically independent, where F has distribution Gamma$(k+1, 1)$. From these facts one can derive that

$$\begin{aligned}\mathbb{E}\big[U_{k0}^{s_0} U_{k1}^{s_1} \cdots U_{kk}^{s_k}\big] &= \mathbb{E}\big[F^{s_+} U_{k0}^{s_0} U_{k1}^{s_1} \cdots U_{kk}^{s_k}\big] \big/ \mathbb{E}(F^{s_+}) \\ &= \mathbb{E}\big[E_0^{s_0} E_1^{s_1} \cdots E_k^{s_k}\big] \big/ \mathbb{E}(F^{s_+}) = \frac{s_0! s_1! \cdots s_k!}{(s_+ + k)!/k!},\end{aligned}$$

so

$$\begin{aligned}T_k(A, \Delta) &= \frac{1}{k!} \sum_{s_0, \ldots, s_k \geq 0} \mathbb{E}\Big[\frac{(U_{k0}A)^{s_0}}{s_0!} B \frac{(U_{k1}A)^{s_1}}{s_1!} \cdots B \frac{(U_{kk}A)^{s_k}}{s_k!}\Big] \\ &= \frac{1}{k!} \mathbb{E}\big[\exp(U_{k0}A) B \exp(U_{k1}A) \cdots B \exp(U_{kk}A)\big].\end{aligned}$$

□

In our proofs of Propositions 5.2 and 5.11 we utilize two elementary bounds for random variables with bounded support. The first one is well-known, but we haven't seen the second one elsewhere.

Lemma 8.3. *Let Y be a random variable with values in $[a, b]$. Then*

$$\mathrm{Var}(Y) \leq (b-a)^2/4 \quad \textit{and} \quad \big|\mathbb{E}\big((Y - \mathbb{E}(Y))^3\big)\big| \leq (b-a)^3/(6\sqrt{3}).$$

In addition we need several properties of an auxiliary function:

Lemma 8.4. *Let $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$ and $M \in \mathbb{R}^{q\times q}_{\mathrm{sym},\geq 0} \setminus \{0\}$. For $t \in \mathbb{R}$ let*

$$g(t) = g(t, A, M) := \log \mathrm{tr}(\exp(-tA)M).$$

This defines a smooth convex function g on $\mathbb{R}$ with the following properties:

$$\begin{aligned}|g'| &\leq \|A\| \quad \textit{with} \quad g'(0) = -\mathrm{tr}(AM)/\mathrm{tr}(M), \\ 0 \leq g'' &\leq \|A\|^2 \quad \textit{with} \quad g''(0) = \mathrm{tr}(A^2M)/\mathrm{tr}(M) - \mathrm{tr}(AM)^2/\mathrm{tr}(M)^2, \\ |g'''| &\leq \|A\|^3 \, 4/\sqrt{27}.\end{aligned}$$

Furthermore, either $g'' > 0$ on $\mathbb{R}$, or there exists an eigenvalue λ of A such that

$$M \in \mathbb{M}(\{x \in \mathbb{R}^q : Ax = \lambda x\}), \quad g' \equiv -\lambda \quad \textit{and} \quad g'' \equiv 0.$$

Proof of Lemma 8.3. It suffices to consider the case $[a, b] = [0, 1]$, because otherwise one could just replace Y with $(Y - a)/(b - a)$. Then

$$\mathrm{Var}(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 \leq \mathbb{E}(Y) - \mathbb{E}(Y)^2 \leq 1/4$$

with equality if, and only if, $Y \in \{0, 1\}$ almost surely and $\mathbb{E}(Y) = 1/2$.

As to the central third moment, with $\mu := \mathbb{E}(Y)$ it suffices to prove that

$$\mathbb{E}((Y-\mu)^3) \;\le\; 1/(6\sqrt{3}), \tag{8.2}$$

because $-(Y-\mu)^3 = ((1-Y)-(1-\mu))^3$. We only have to consider the situation that $0 < \mu < 1$ with strictly positive probabilities $p_0 := \mathbb{P}(Y < \mu)$ and $p_1 := \mathbb{P}(Y \ge \mu)$, because otherwise $Y = \mu$ almost surely. Note that $h(x) := (x-\mu)^3$ is concave on $[0,\mu]$ and convex on $[\mu,1]$. Hence with

$$x_0 \;:=\; \mathbb{E}(Y \,|\, Y < \mu) \quad\text{and}\quad x_1 \;:=\; \mathbb{E}(Y \,|\, Y \ge \mu)$$

we may conclude from Jensen's inequality that

$$\begin{aligned}
\mathbb{E}((Y-\mu)^3) \;&=\; p_0\mathbb{E}(h(Y) \,|\, Y < \mu) + p_1\mathbb{E}(h(Y) \,|\, Y \ge \mu) \\
&\le\; p_0(x_0-\mu)^3 + p_1\mathbb{E}(h(Y) \,|\, Y \ge \mu) \\
&\le\; p_0(x_0-\mu)^3 + p_1\mathbb{E}\Big(\frac{1-Y}{1-\mu}h(\mu) + \frac{Y-\mu}{1-\mu}h(1) \,\Big|\, Y \ge \mu\Big) \\
&=\; p_0(x_0-\mu)^3 + p_1\mathbb{E}\Big(\frac{Y-\mu}{1-\mu}(1-\mu)^3 \,\Big|\, Y \ge \mu\Big) \\
&=\; p_0(x_0-\mu)^3 + p_1(x_1-\mu)(1-\mu)^2.
\end{aligned}$$

Equality holds if

$$Y \;\sim\; p_0\delta_{x_0} + \frac{p_1(1-x_1)}{1-\mu}\delta_\mu + \frac{p_1(x_1-\mu)}{1-\mu}\delta_1.$$

Note that in the latter case, $\mathbb{E}(Y)$ is still equal to μ, because $p_0x_0 + p_1x_1 = \mu$. If we replace $\mathcal{L}(Y)$ with $\mathcal{L}(Y \,|\, Y \ne \mu)$, the mean does not change, but $\mathbb{E}((Y-\mu)^3)$ increases by the factor $1/\mathbb{P}(Y \ne \mu)$. Thus it even suffices to consider distributions $\mathcal{L}(Y)$ which are concentrated on two points $x_0 \in [0,1)$ and 1. Finally, in case of $x_0 > 0$ we could replace Y and μ with $(Y-x_0)/(1-x_0)$ and $(\mu-x_0)/(1-x_0) = \mathbb{P}(Y=1)$, respectively. This would increase $\mathbb{E}((Y-\mu)^3)$ by a factor $(1-x_0)^{-3}$ and lead to a random variable with values in $\{0,1\}$.

Finally we have to maximize

$$(1-\mu)(0-\mu)^3 + \mu(1-\mu)^3 \;=\; \mu(1-\mu)(1-2\mu)$$

over all $\mu \in (0,1)$. With $u := 1-2\mu \in (-1,1)$ one may write

$$\mu(1-\mu)(1-2\mu) \;=\; 4^{-1}(1-u^2)u \;\le\; 1/(6\sqrt{3})$$

with equality for $u = 1/\sqrt{3}$. □

Proof of Lemma 8.4. Let $A = \sum_{i=1}^q \lambda_i(A)u_iu_i^\top$ with an orthonormal basis $u_1, u_2, \ldots, u_q$ of $\mathbb{R}^q$. Then $\mathrm{tr}(M) = \sum_{i=1}^q u_i^\top M u_i$ and

$$g(t) \;=\; \log\Big(\sum_{i=1}^q e^{-t\lambda_i(A)}u_i^\top M u_i\Big) \;=\; \log\mathrm{tr}(M) + \log\mathbb{E}(e^{tY}),$$

where $Y \sim \sum_{i=1}^{q} p_i \delta_{-\lambda_i(A)}$ with $p_i := u_i^\top M u_i / \operatorname{tr}(M)$. Elementary calculations show that

$$\begin{aligned} g'(t) &= \mathbb{E}(e^{tY}Y)/\mathbb{E}(e^{tY}), \\ g''(t) &= \mathbb{E}(e^{tY}Y^2)/\mathbb{E}(e^{tY}) - \mathbb{E}(e^{tY}Y)^2/\mathbb{E}(e^{tY})^2, \\ g'''(t) &= \mathbb{E}(e^{tY}Y^3)/\mathbb{E}(e^{tY}) - 3\mathbb{E}(e^{tY}Y^2)\mathbb{E}(e^{tY}Y)/\mathbb{E}(e^{tY})^2 \\ &\quad + 2\mathbb{E}(e^{tY}Y)^3/\mathbb{E}(e^{tY})^3. \end{aligned}$$

Defining the modified distribution $\mathbb{P}_t$ via $\mathbb{P}_t(B) := \mathbb{E}(e^{tY}1_B)/\mathbb{E}(e^{tY})$, we may rewrite this as

$$g'(t) = \mathbb{E}_t(Y), \quad g''(t) = \operatorname{Var}_t(Y) \quad \text{and} \quad g'''(t) = \mathbb{E}_t\big((Y - \mathbb{E}_t(Y))^3\big).$$

In particular, $g'(0) = \mathbb{E}(Y)$ equals $-\operatorname{tr}(AM)/\operatorname{tr}(M)$, and $g''(0) = \operatorname{Var}(Y)$ equals $\operatorname{tr}(A^2M)/\operatorname{tr}(M) - \operatorname{tr}(AM)^2/\operatorname{tr}(M)^2$.

Note that $|Y| \le \|A\|$, so $|g'| \le \|A\|$. Further it follows from Lemma 8.3 with $[a,b] = \big[-\|A\|, \|A\|\big]$ that $0 \le g''(0) \le \|A\|^2$, and $|g'''| \le \|A\|^3 4/\sqrt{27}$.

Finally, for any $t_o \in \mathbb{R}$ the equation $g''(t_o) = 0$ is equivalent to Y being constant almost surely with respect to $\mathbb{P}_{t_o}$. But this means that for some eigenvalue λ of A,

$$u_i^\top M u_i = 0 \quad \text{whenever} \quad \lambda_i(A) \ne \lambda,$$

so $M \in \mathbb{M}(\{x \in \mathbb{R}^q : Ax = \lambda x\})$. This implies that $g(t) = g(0) - \lambda t$ for all $t \in \mathbb{R}$, whence $g' \equiv -\lambda$ and $g'' \equiv 0$. □

Proof of Proposition 5.2. Note first that

$$L_\rho(\exp(A), Q) = \operatorname{tr}(A) + \int \big[\rho(\operatorname{tr}(\exp(-A)M)) - \rho(\operatorname{tr}(M))\big]\, Q(dM).$$

For fixed $M \in \mathbb{R}^{q\times q}_{\mathrm{sym},\ge 0} \setminus \{0\}$ let $a := \operatorname{tr}(M) > 0$ and $b := \operatorname{tr}(\exp(-A)M)$. Then $b/a \in \big[\lambda_{\min}(\exp(-A)), \lambda_{\max}(\exp(-A))\big] \subset [e^{-\|A\|}, e^{\|A\|}]$ by Lemma 4.7. Hence Lemma 4.8 implies that $\rho(\operatorname{tr}(\exp(-A)M)) - \rho(\operatorname{tr}(M))$ equals

$$\rho(b) - \rho(a) = \psi(a)\log(b/a) + r_1(a,b)$$

with

$$\begin{aligned} |r_1(a,b)| &\le \big(\psi(\max\{a,b\}) - \psi(\min\{a,b\})\big)\,|\log(b/a)| \\ &\le \big(\psi\big(e^{\|A\|}\operatorname{tr}(M)\big) - \psi\big(e^{-\|A\|}\operatorname{tr}(M)\big)\big)\|A\|. \end{aligned}$$

Moreover, $\log(b/a) = g(1) - g(0)$ with $g = g(\cdot, A, M)$ as in Lemma 8.4. Hence for a suitable number $\xi \in (0,1)$,

$$g(1) - g(0) = g'(0) + g''(\xi)/2$$

where $g'(0) = -\operatorname{tr}(AM)/\operatorname{tr}(M)$ and $0 \le g''(\xi) \le \|A\|^2$. All in all we obtain the expansion

$$\begin{aligned}\rho(b) - \rho(a) &= \psi(a)g'(0) + \psi(a)g''(\xi)/2 + r_1(a,b) \\ &= -\rho'(\operatorname{tr}(M))\operatorname{tr}(AM) + \psi(\operatorname{tr}(M))g''(\xi)/2 + r_1(a,b).\end{aligned}$$

Consequently

$$L_\rho(\exp(A), Q) = \operatorname{tr}(A) - \int \rho'(\operatorname{tr}(M))\operatorname{tr}(AM)\,Q(dM) + R_\rho(A,Q),$$

where

$$|R_\rho(A,Q)| \le \big(J_\rho(e^{\|A\|}, Q) - J_\rho(e^{-\|A\|}, Q)\big)\|A\| + J_\rho(Q)\|A\|^2/2.$$

Moreover,

$$\operatorname{tr}(A) - \int \rho'(\operatorname{tr}(M))\operatorname{tr}(AM)\,Q(dM) = \langle A, G_\rho(Q)\rangle$$

with $G_\rho(Q) = I_q - \int \rho'(\operatorname{tr}(M))M\,Q(dM) = I_q - \Psi_\rho(Q)$, and the inequalities $|\operatorname{tr}(A)| \le q\|A\|$ and $|\operatorname{tr}(AM)\| \le \|A\|\operatorname{tr}(M)$ imply that $\big|\langle A, G_\rho(Q)\rangle\big|$ is bounded by $(q + J_\rho(Q))\|A\|$. □

Proof of Corollary 5.3. For fixed $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ let $B := \Sigma^{1/2}$. If $\Delta \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$ with $\|\Delta\| < \lambda_{\min}(\Sigma)$, then $\Sigma + \Delta \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$, too, and we may write

$$\Sigma + \Delta = B(I_q + B^{-1}\Delta B^{-1})B = B\exp(A(\Delta))B$$

with $A(\Delta) := \log(I_q + B^{-1}\Delta B^{-1})$, whence

$$L_\rho(\Sigma + \Delta, Q) - L_\rho(\Sigma, Q) = L_\rho(\exp(A(\Delta)), Q_B).$$

As $\Delta \to 0$,

$$A(\Delta) = B^{-1}\Delta B^{-1} + O(\|\Delta\|^2),$$

so it follows from Proposition 5.2 that

$$\begin{aligned}L_\rho(\exp(A(\Delta)), Q_B) &= \langle B^{-1}\Delta B^{-1}, G_\rho(Q_B)\rangle + o(\|\Delta\|) \\ &= \langle \Delta, B^{-1}G_\rho(Q_B)B^{-1}\rangle + o(\|\Delta\|).\end{aligned}$$

Consequently, $\nabla L_\rho(\Sigma, Q)$ equals

$$B^{-1}G_\rho(Q_B)B^{-1} = \Sigma^{-1} - \int \rho'(\operatorname{tr}(\Sigma^{-1}M))\Sigma^{-1}M\Sigma^{-1}\,Q(dM).$$

By dominated convergence, this is continuous in Σ, because $\Sigma \mapsto \Sigma^{-1}$ is continuous, ρ' is continuous on $(0,\infty)$, and the norm of the integrand on the right hand side is not greater than $\lambda_{\min}(\Sigma)^{-1}\psi(\lambda_{\min}(\Sigma)^{-1}\operatorname{tr}(M))$.

For a compact convex set $K \subset \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ and $\Sigma_0, \Sigma_1 \in K$ define the convex combination $\Sigma_t := (1-t)\Sigma_0 + t\Sigma_1$ for $t \in [0,1]$. Then $L_\rho(\Sigma_t, Q)$ is differentiable in t with derivative $\langle \Sigma_1 - \Sigma_0, \nabla L_\rho(\Sigma_t, Q)\rangle$. Hence for a suitable point $\xi \in (0,1)$ and $B := \Sigma_\xi^{1/2}$ it follows from the bounds in Proposition 5.2 and inequality (5.1) that

$$\begin{aligned}
\big|L_\rho(\Sigma_1, Q) - L_\rho(\Sigma_0)\big| &= \big|\langle \Sigma_1 - \Sigma_0, \nabla L_\rho(\Sigma_\xi, Q)\rangle\big| \\
&= \big|\big\langle B^{-1}(\Sigma_1 - \Sigma_0)B^{-1}, G_\rho(Q_B)\big\rangle\big| \\
&\le (q + J_\rho(Q_B))\big\|B^{-1}(\Sigma_1 - \Sigma_0)B^{-1}\big\| \\
&\le \big(q + J_\rho(\lambda_{\min}(\Sigma_\xi)^{-1}, Q)\big)\lambda_{\min}(\Sigma)^{-1}\big\|\Sigma_1 - \Sigma_0\big\| \\
&\le (q + J_\rho(\Lambda_K, Q))\Lambda_K\big\|\Sigma_1 - \Sigma_0\big\|.
\end{aligned}$$

□

Proof of Proposition 5.4. Note first that by (4.4),

$$\begin{aligned}
&L_\rho(B\exp(tA)B^\top, Q) - L_\rho(BB^\top, Q) \\
&\quad = L_\rho(\exp(tA), Q_B) \\
&\quad = t\cdot \mathrm{tr}(A) + \int \big[\rho(\mathrm{tr}(\exp(-tA)M)) - \rho(\mathrm{tr}(M))\big]\, Q(dM).
\end{aligned}$$

Thus we consider a fixed matrix $M \in \mathbb{R}^{q\times q}_{\mathrm{sym},\ge 0} \setminus \{0\}$ and verify convexity of

$$h(t) = h(t, A, M) \;:=\; \rho(e^{g(t)})$$

with $g(t) = \log \mathrm{tr}(\exp(-tA)M))$ as in Lemma 8.4. Indeed,

$$h'(t) \;=\; \rho'(e^{g(t)})e^{g(t)}g'(t) \;=\; \psi(e^{g(t)})g'(t)$$

is monotone increasing in $t \in \mathbb{R}$. For if $s < t$, then

$$\begin{aligned}
&\psi(e^{g(t)})g'(t) - \psi(e^{g(s)})g'(s) \\
&\qquad = \begin{cases} \big(\psi(e^{g(t)}) - \psi(e^{g(s)})\big)g'(s) + \psi(e^{g(t)})\big(g'(t) - g'(s)\big) \\ \big(\psi(e^{g(t)}) - \psi(e^{g(s)})\big)g'(t) + \psi(e^{g(s)})\big(g'(t) - g'(s)\big) \end{cases} \\
&\qquad \ge \begin{cases} \big(\psi(e^{g(t)}) - \psi(e^{g(s)})\big)g'(s) \\ \big(\psi(e^{g(t)}) - \psi(e^{g(s)})\big)g'(t) \end{cases} \qquad (8.3) \\
&\qquad \ge 0. \qquad (8.4)
\end{aligned}$$

Inequality (8.3) follows from ψ being positive and g' being non-decreasing. Inequality (8.4) follows from ψ being non-decreasing and g being convex. For if $\psi(e^{g(t)}) - \psi(e^{g(s)}) > 0$, then $g(t) - g(s) > 0$ and thus $g'(t) > 0$. Likewise $\psi(e^{g(t)}) - \psi(e^{g(s)}) < 0$ implies that $g(t) - g(s) < 0$ whence $g'(s) < 0$.

Concerning strict convexity, recall from Lemma 8.4 that either $g'' > 0$ on $\mathbb{R}$, or $g'' \equiv 0$ and $M \in \bigcup_{i=1}^{\ell} \mathbb{M}(\mathbb{V}_i)$. Hence, in Case 0, $\psi(e^g)g' = qg'$ is strictly increasing if, and only if, $M \notin \bigcup_{i=1}^{\ell} \mathbb{M}(\mathbb{V}_i)$. Consequently, $t \mapsto L_\rho(B\exp(tA)B^\top, Q)$ is strictly convex if, and only if, $Q_B\big(\bigcup_{i=1}^{\ell} \mathbb{M}(\mathbb{V}_i)\big) = Q\big(\bigcup_{i=1}^{\ell} \mathbb{M}(B\mathbb{V}_i)\big) < 1$.

In Case 1, inequality (8.3) is strict, unless $g'' \equiv 0$. But in the latter case, $g(t) = g(0)+g'(0)t$ and $g'(t) = g'(0)$, so inequality (8.4) is strict, unless $g'(0) = 0$. Hence h is strictly convex unless g is constant. But this is equivalent to saying that $M \in \mathbb{M}(\mathbb{V}_0)$. Consequently, $t \mapsto L_\rho(B\exp(tA)B^\top, Q)$ is strictly convex, unless $Q_B(\mathbb{M}(\mathbb{V}_0)) = Q(\mathbb{M}(B\mathbb{V}_0)) = 1$. □

Proof of Proposition 5.5. Since Conditions 0 and 1 are not affected by replacing Q with Q_B, we may restrict our attention to $B = I_q$. Let $\mathbb{W} := \{A \in \mathbb{R}^{q\times q}_{\mathrm{sym}} : \mathrm{tr}(A) = 0\}$ in Case 0 and $\mathbb{W} := \mathbb{R}^{q\times q}_{\mathrm{sym}}$ in Case 1. For $A \in \mathbb{W}$ and $t \in \mathbb{R}$ let

$$h(t, A) \;:=\; L_\rho(\exp(tA), Q).$$

We know from Proposition 5.4 that h is convex in the first argument. Moreover, the derivative $h'(t, A) = \partial h(t, A)/\partial t$ is given by

$$h'(t, A) \;=\; \mathrm{tr}(A) + \int \rho'\big(\mathrm{tr}(\exp(-tA)M)\big)\,\mathrm{tr}(-A\exp(-tA)M)\,Q(dM).$$

This could be verified directly or derived from Proposition 5.2, because $h(t+s, A) - h(t, A) = L_\rho(\exp(sA), Q_{\exp(tA/2)})$. The derivative $h'(t, A)$ is continuous in A, which implies the following equivalence:

$$\lim_{\|B\|\to\infty, B\in\mathbb{W}} L_\rho(\exp(B), Q) \;=\; \infty \tag{8.5}$$

if, and only if,

$$h'(A) := \lim_{t\to\infty} h'(t, A) \;>\; 0 \quad \text{for any fixed } A \in \mathbb{W}\setminus\{0\}. \tag{8.6}$$

To see this, note first that $h'(A) \le 0$ is equivalent to $h(\cdot, A)$ being non-increasing. Thus a violation of (8.6) would imply a violation of (8.5). Now suppose that (8.6) holds true. Since $h'(t, A)$ is non-decreasing in $t \ge 0$ and continuous in $A \in \mathbb{S}(\mathbb{W}) := \{A \in \mathbb{W} : \|A\| = 1\}$,

$$U(t) \;:=\; \big\{A \in \mathbb{S}(\mathbb{W}) : h'(t, A) > 0\big\}$$

is an open subset of $\mathbb{S}(\mathbb{W})$ with $U(s) \subset U(t)$ whenever $s < t$. Moreover, (8.6) entails that $\bigcup_{t\ge0} U(t) = \mathbb{S}(\mathbb{W})$. But the latter set is compact, so $U(t_o) = \mathbb{S}(\mathbb{W})$ for some $t_o \ge 0$. Now for $t \ge t_o$ we have by the convexity of h in the first argument,

$$\begin{aligned}
\min_{B\in\mathbb{W}\,:\,\|B\|=t} L_\rho(\exp(B), Q) \;&=\; \min_{A\in\mathbb{S}(\mathbb{W})} h(t, A) \\
&\ge\; \min_{A\in\mathbb{S}(\mathbb{W})} h(t_o, A) + (t - t_o) \min_{A\in\mathbb{S}(\mathbb{W})} h'(t_o, A) \\
&\to\; \infty \quad \text{as } t \to \infty,
\end{aligned}$$

i.e. (8.5) is satisfied, too.

Now we determine the limit $h'(A)$ for fixed $A \in \mathbb{W}\setminus\{0\}$. To this end we write $A = -\sum_{i=1}^q \beta_i u_i u_i^\top$ with $\beta_i := -\lambda_i(A)$ and an orthonormal basis $u_1, u_2, \ldots, u_q$ of $\mathbb{R}^q$. Then

$$h'(t,A) \;=\; -\sum_{i=1}^q \beta_i + \int \psi\Big(\sum_{i=1}^q u_i^\top M u_i\, e^{t\beta_i}\Big) \frac{\sum_{i=1}^q \beta_i u_i^\top M u_i\, e^{t\beta_i}}{\sum_{i=1}^q u_i^\top M u_i\, e^{t\beta_i}}\, Q(dM)$$

with $\psi(0)\cdot 0/0 := 0$. As shown in the proof of Proposition 5.4, the integrand on the right hand side is non-decreasing in $t \ge 0$. Let $\mathbb{V}_0 := \{0\}$ and $\mathbb{V}_j := \operatorname{span}(u_1,\ldots,u_j)$ for $1 \le j \le q$. If $M \in \mathbb{M}(\mathbb{V}_j)\setminus\mathbb{M}(\mathbb{V}_{j-1})$, then $u_j^\top M u_j > 0 = u_k^\top M u_k$ for $j < k \le q$, and one can easily derive from $\beta_1 \le \beta_2 \le \cdots \le \beta_q$ that

$$\lim_{t\to\infty} \psi\Big(\sum_{i=1}^q u_i^\top M u_i\, e^{t\beta_i}\Big) \frac{\sum_{i=1}^q \beta_i u_i^\top M u_i\, e^{t\beta_i}}{\sum_{i=1}^q u_i^\top M u_i\, e^{t\beta_i}} \;=\; \begin{cases} q\beta_j & \text{in Case 0} \\ \psi(\infty)\beta_j^+ & \text{in Case 1} \end{cases}$$

with the usual notation $a^\pm = \max(\pm a, 0)$ for real numbers a. Thus it follows from monotone convergence that

$$h'(A) \;=\; -\sum_{i=1}^q \beta_i + \begin{cases} q\sum_{j=1}^q \beta_j Q\big(\mathbb{M}(\mathbb{V}_j)\setminus\mathbb{M}(\mathbb{V}_{j-1})\big) & \text{in Case 0,} \\ \psi(\infty)\sum_{j=1}^q \beta_j^+ Q\big(\mathbb{M}(\mathbb{V}_j)\setminus\mathbb{M}(\mathbb{V}_{j-1})\big) & \text{in Case 1.} \end{cases}$$

In Case 0, define $\gamma_d := \beta_{d+1} - \beta_d$ for $d = 1,\ldots,q-1$. Then

$$\begin{aligned} h'(A) \;&=\; q\sum_{j=1}^q \beta_j\big[-1/q + Q(\mathbb{M}(\mathbb{V}_j)) - Q(\mathbb{M}(\mathbb{V}_{j-1}))\big] \\ &=\; q\sum_{j=1}^q \beta_j\big[Q(\mathbb{M}(\mathbb{V}_j)) - j/q - Q(\mathbb{M}(\mathbb{V}_{j-1})) + (j-1)/q\big] \\ &=\; q\sum_{j=1}^{q-1} \beta_j\big[Q(\mathbb{M}(\mathbb{V}_j)) - j/q\big] + q\sum_{j=2}^q \beta_j\big[(j-1)/q - Q(\mathbb{M}(\mathbb{V}_{j-1}))\big] \\ &=\; q\sum_{d=1}^{q-1} \gamma_d\big[d/q - Q(\mathbb{M}(\mathbb{V}_d))\big], \end{aligned}$$

where we utilized that $Q(\mathbb{M}(\mathbb{V}_0)) = Q(\{0\}) = 0$ and $Q(\mathbb{M}(\mathbb{V}_q)) = Q(\mathbb{R}^{q\times q}_{\mathrm{sym}}) = 1$. Since all γ_d are non-negative with $\sum_{d=1}^{q-1}\gamma_d = \beta_q - \beta_1 > 0$, Condition 0 implies clearly that $h'(A) > 0$. On the other hand, if $Q(\mathbb{M}(\mathbb{V})) \ge j/q$ for some $\mathbb{V} \in \mathcal{V}_q$ with $d := \dim(\mathbb{V}) \in [1,q)$, we may choose the basis $u_1, u_2, \ldots, u_q$ such that $\mathbb{V} = \mathbb{V}_d$, and with $\beta_i := 1_{[i>d]} - (q-d)/q$, the matrix $A = -\sum_{i=1}^q \beta_i u_i u_i^\top$ satisfies $h'(A) = q\big[d/q - Q(\mathbb{M}(\mathbb{V}_d))\big] \le 0$. Consequently, (8.6) and Condition 0 are equivalent in Case 0.

In Case 1, let $\gamma_d := \beta_{d+1}^+ - \beta_d^+$ for $d = 0, 1, \ldots, q-1$, where $\beta_0^+ := 0$. Then $-\sum_{i=1}^q \beta_i$ is equal to

$$\sum_{i=1}^q \beta_i^- - \sum_{i=1}^q (\beta_i^+ - \beta_0^+) \;=\; \sum_{i=1}^q \beta_i^- - \sum_{i=1}^q \sum_{d=0}^{i-1} \gamma_d \;=\; \sum_{i=1}^q \beta_i^- - \sum_{d=0}^{q-1} \gamma_d (q-d)$$

and $\sum_{j=1}^q \beta_j^+ Q\big(\mathbb{M}(\mathbb{V}_j) \setminus \mathbb{M}(\mathbb{V}_{j-1})\big)$ may be written as

$$\sum_{j=1}^q \beta_j^+ \big[1 - Q(\mathbb{M}(\mathbb{V}_{j-1}))\big] - \sum_{j=0}^{q-1} \beta_j^+ \big[1 - Q(\mathbb{M}(\mathbb{V}_j))\big] \;=\; \sum_{d=0}^{q-1} \gamma_d \big[1 - Q(\mathbb{M}(\mathbb{V}_d))\big].$$

Consequently,

$$h'(A) \;=\; \sum_{i=1}^q \beta_i^- + \sum_{d=0}^{q-1} \gamma_d \Big(\psi(\infty) \big[1 - Q(\mathbb{M}(\mathbb{V}_d))\big] - (q-d) \Big).$$

Again one can easily deduce from $\gamma_d \geq 0$ and $\sum_{d=0}^{q-1} \gamma_d = \beta_q^+ = \max_i \beta_i^+$ that Condition 1 implies (8.6). On the other hand, if $Q(\mathbb{M}(\mathbb{V})) \geq 1 - (q-d)/\psi(\infty)$ for some $\mathbb{V} \in \mathcal{V}_q$ with $d := \dim(\mathbb{V}) \in [0, q)$, we may choose the basis $u_1, u_2, \ldots, u_q$ such that $\mathbb{V} = \mathbb{V}_d$, and with $\beta_i := 1_{[i>d]}$ we obtain a matrix A such that $h'(A) \leq 0$. Consequently, (8.6) and Condition 1 are equivalent in Case 1. □

Proof of Lemma 5.10. Suppose that Condition (5.5) is satisfied; in other words,

$$\partial \log \phi(t)/\partial t \;\leq\; \kappa t^{-1} \quad \text{for all } t > 0.$$

Now fix arbitrary $s > 0$ and $\lambda > 1$. For any integer $\ell > 1$,

$$\begin{aligned} \log \phi(\lambda s) - \log \phi(s) \;&=\; \sum_{i=1}^{\ell} \big(\log \phi(\lambda^{i/\ell} s) - \log \phi(\lambda^{(i-1)/\ell} s) \big) \\ &\leq\; \sum_{i=1}^{\ell} (\lambda^{i/\ell} s - \lambda^{(i-1)/\ell} s) \kappa (\lambda^{(i-1)/\ell} s)^{-1} \\ &=\; \kappa \ell (\lambda^{1/\ell} - 1) \;\to\; \kappa \log \lambda \quad \text{as } \ell \to \infty. \end{aligned}$$

Consequently, $\log \phi(\lambda s) - \log \phi(s) \leq \kappa \log \lambda$, which proves Condition (5.6).

On the other hand, if Condition (5.6) is satisfied, then for $s > 0$,

$$s\phi'(s) \;=\; \lim_{\lambda \downarrow 1} \frac{\phi(\lambda s) - \phi(s)}{\lambda - 1} \;\leq\; \lim_{\lambda \downarrow 1} \frac{(\lambda^\kappa - 1)\phi(s)}{\lambda - 1} \;=\; \kappa \phi(s).$$

Hence Condition (5.5) is satisfied as well. □

Proof of Proposition 5.11. As in the proof of Proposition 5.2 we start from

$$L_\rho(\exp(A), Q) \;=\; \operatorname{tr}(A) + \int \big[\rho(\operatorname{tr}(\exp(-A)M)) - \rho(\operatorname{tr}(M)) \big] \, Q(dM)$$

and analyze for a fixed $M \in \mathbb{R}^{q \times q}_{\mathrm{sym},\geq 0} \setminus \{0\}$ the difference $\rho(b) - \rho(a)$, where $a := \operatorname{tr}(M) > 0$ and $b := \operatorname{tr}(\exp(-A)M)$.

Recall first that $b/a \in [e^{-\|A\|}, e^{\|A\|}]$. For $x \in \mathbb{R}$ define $f(x) := \rho(e^x a)$. Then $f'(x) = \rho'(e^x a)e^x a = \psi(e^x a)$, and $f''(x) = \psi'(e^x a)e^x a = \psi_2(e^x a)$. Consequently, for a suitable point ξ between 0 and $\log(b/a)$,

$$\begin{aligned}\rho(b) - \rho(a) &= f(\log(b/a)) - f(0) \\ &= \psi(a)\log(b/a) + \psi_2(e^\xi a)\log(b/a)^2/2 \\ &= \psi(a)\log(b/a) + \psi_2(a)\log(b/a)^2/2 + r_2(a,b),\end{aligned}$$

where

$$|r_2(a,b)| \le \sup_{z\in[-\|A\|,\|A\|]} \big|\psi_2(e^z s) - \psi_2(s)\big| \|A\|^2/2.$$

Now we utilize the fact that $\log(b/a) = g(1) - g(0)$ with the auxiliary function $g(t) := \log \operatorname{tr}(\exp(-tA)M)$ from Lemma 8.4. In particular, $|g(1) - g(0) - g'(0)| \le \|A\|^2/2$ and $|g(1) - g(0) - g'(0) - g''(0)/2| \le \|A\|^3(4/\sqrt{27})/6 \le \|A\|^3/7$. Consequently,

$$\begin{aligned}\psi(a)\log(b/a) &= \psi(a)\big(g'(0) + g''(0)/2\big) + r_3(a,b), \\ \psi_2(a)\log(b/a)^2/2 &= \psi_2(a)g'(0)^2/2 + r_4(a,b),\end{aligned}$$

where

$$\begin{aligned}|r_3(a,b)| &< \psi(a)\|A\|^3/7, \\ |r_4(a,b)| &\le \psi_2(a)\big|\log(b/a)^2 - g'(0)^2\big|/2 \\ &\le \kappa\psi(a)\big|\log(b/a) - g'(0)\big|\big(|\log(b/a)| + |g'(0)|\big)/2 \\ &\le \kappa\psi(a)\|A\|^3.\end{aligned}$$

All in all this shows that

$$\rho(b) - \rho(a) = \psi(a)g'(0) + \psi(a)g''(0)/2 + \psi_2(a)g'(0)^2/2 + r_*(a,b)$$

with

$$|r_*(a,b)| \le \sup_{z\in[-\|A\|,\|A\|]} \big|\psi_2(e^z a) - \psi_2(a)\big|\|A\|^2/2 + \psi(a)(\kappa + 1/7)\|A\|^3.$$

Note that $\psi(a)g'(0) = \rho'(\operatorname{tr}(M))\operatorname{tr}(AM)$. Moreover, it follows from $|g'| \le \|A\|$, $0 \le g'' \le \|A\|^2$ and $\psi, \psi_2 \ge 0$ that

$$\begin{aligned}0 \le \psi(a)g''(0) + \psi_2(a)g'(0)^2 &\le \psi(\operatorname{tr}(M))\|A\|^2 + \psi_2(\operatorname{tr}(M))\|A\|^2 \\ &\le (1+\kappa)\psi(\operatorname{tr}(M))\|A\|^2.\end{aligned}$$

Furthermore, elementary calculations show that

$$\psi(a)g''(0) + \psi_2(a)g'(0)^2 = \rho'(\operatorname{tr}(M))\operatorname{tr}(A^2M) + \rho''(\operatorname{tr}(M))\operatorname{tr}(AM)^2.$$

Consequently,

$$L(\exp(A), Q) = \langle A, G_\rho(A)\rangle + 2^{-1}H_\rho(A,Q) + R_{\rho,2}(A,Q)$$

with the quadratic term

$$H_\rho(A,Q) \;=\; \int \big(\rho'(\mathrm{tr}(M))\,\mathrm{tr}(A^2M) + \rho''(\mathrm{tr}(M))\,\mathrm{tr}(AM)^2\big)\,Q(dM)$$

and a remainder $R_{\rho,2}(A,Q)$ satisfying the asserted bounds (5.8) and (5.9).

It remains to prove inequality (5.10). Since $H_\rho(A,Q)$ is the integral of the term $\psi(a)g''(0) + \psi_2(a)g'(0)^2 \geq 0$ with $a = \mathrm{tr}(M)$ and $g = g(\cdot, A, M)$, it is equal to 0 if, and only if, $\psi(a)g''(0) + \psi_2(a)g'(0)^2$ for Q-almost all M. Based on Lemma 8.4 we may argue as follows: In Case 0, $\psi(a)g''(0)+\psi_2(a)g'(0)^2 = qg''(0)$ equals zero if, and only if, $M \in \bigcup_{i=1}^\ell \mathbb{M}(\mathbb{V}_i)$. Hence $H_\rho(A,Q) > 0$ is equivalent to $Q\big(\bigcup_{i=1}^\ell \mathbb{M}(\mathbb{V}_i) < 1$. In Case 1, both $\psi(a)$ and $\psi_2(a)$ are strictly positive while $g''(0) \geq 0$. Hence $\psi(a)g''(0) + \psi_2(a)g'(0)^2$ equals zero if, and only if, $g''(0) = g'(0) = 0$, which is equivalent to $M \in \mathbb{M}(\mathbb{V}_0)$. Consequently, $H_\rho(A,Q) > 0$ if, and only if, $Q(\mathbb{M}(\mathbb{V}_0)) < 1$. □

8.4. Proofs for Section 6

In the proof of Theorem 6.3 we utilize a well-known elementary fact about weak convergence, adapted to random distributions:

Lemma 8.5. *Let Q be a fixed and $\widehat{Q}_1, \widehat{Q}_2, \widehat{Q}_3, \ldots$ be random probability distributions on a metric space $(\mathbb{Y}, d)$ with the following two properties: For any bounded and continuous function $f : \mathbb{Y} \to \mathbb{R}$,*

$$\int f\,d\widehat{Q}_n \;\to_p\; \int f\,dQ.$$

Further, for a particular continuous function $\phi : \mathbb{Y} \to [0,\infty)$, $\int \phi\,d\widehat{Q}_n < \infty$ almost surely for all n, and

$$\int \phi\,d\widehat{Q}_n \;\to_p\; \int \phi\,dQ < \infty.$$

Then

$$\int f\,d\widehat{Q}_n \;\to_p\; \int f\,dQ$$

for any continuous function $f : \mathbb{Y} \to \mathbb{R}$ such that $|f|/(1+\phi)$ is bounded on $\mathbb{Y}$.

Proof of Lemma 8.5. It suffices to consider any continuous function $f : \mathbb{Y} \to \mathbb{R}$ such that $|f| \leq \tilde{\phi} := 1 + \phi$. For any fixed number $R \geq 1$ let

$$f_R(y) \;:=\; \mathrm{sign}(f(y))\,\min\{|f(y)|, R\}.$$

Then

$$\begin{aligned}\Big|\int f\,d\widehat{Q}_n - \int f\,dQ\Big| \;&\leq\; \int |f - f_R|\,d\widehat{Q}_n + \int |f - f_R|\,dQ \\ &\quad+\; \Big|\int f_R\,d\widehat{Q}_n - \int f_R\,dQ\Big| \\ &=\; \int |f - f_R|\,d\widehat{Q}_n + \int |f - f_R|\,dQ + o_p(1)\end{aligned}$$

by our first assumption. But $|f - f_R| = (|f| - R)^+ \le (\tilde{\phi} - R)^+ = (\phi - R + 1)^+$, so

$$\begin{aligned}\int |f - f_R|\, d\widehat{Q}_n &\le \int (\phi - R + 1)^+\, d\widehat{Q}_n \\ &= \int \phi\, d\widehat{Q}_n - \int \min\{\phi, R-1\}\, d\widehat{Q}_n \\ &\to_p \int \phi\, dQ - \int \min\{\phi, R-1\}\, dQ = \int (\phi - R + 1)^+\, dQ\end{aligned}$$

by our assumptions. Consequently,

$$\Big|\int f\, d\widehat{Q}_n - \int f\, dQ\Big| \le 2\int (\phi - R + 1)^+\, dQ + o_p(1),$$

and the integral on the right hand is arbitrarily small for sufficiently large R. □

Proof of Theorem 6.3. By linear equivariance we may assume without loss of generality that $\mathbf{\Sigma}_\rho(Q) = I_q$. Let $\mathbb{W} := \{A \in \mathbb{R}^{q\times q}_{\mathrm{sym}} : \mathrm{tr}(A) = 0\}$ in Case 0, and $\mathbb{W} := \mathbb{R}^{q\times q}_{\mathrm{sym}}$ in Case 1. For any fixed $\delta > 0$, the set $K_\delta := \{A \in \mathbb{W} : \|A\| \le \delta\}$ is compact, and for $A \in K_\delta$,

$$f(A, M) := \mathrm{tr}(A) + \big[\rho\big(\mathrm{tr}(\exp(-A)M)\big) - \rho(\mathrm{tr}(M))\big]$$

is continuous in $M \in \mathbb{Y}$ with

$$|f(A, M)| \le q\delta + \psi(e^\delta \mathrm{tr}(M))\delta$$

by Lemmas 4.7 and 4.8. If δ is sufficiently small, $\psi(e^\delta \mathrm{tr}(M)) \le \psi(\mathrm{tr}(\Sigma_o^{-1}M))$ for any $M \in \mathbb{Y}$. Then it follows from Lemma 8.5 that

$$\begin{aligned}L_\rho(\exp(A), Q_n) &= \int f(A, M), \widehat{Q}_n(dM) \\ &\to_p \int f(A, M)\, Q(dM) = L_\rho(\exp(A), Q)\end{aligned}$$

for any fixed $A \in K_\delta$. Moreover it follows from Corollary 5.3 and the first part of Lemma 5.1 that

$$\begin{aligned}\big|L_\rho(\exp(A), \widehat{Q}_n) - L_\rho(\exp(B), \widehat{Q}_n)\big| &\le J(e^\delta, \widehat{Q}_n)e^\delta \|\exp(A) - \exp(B)\| \\ &\le J(e^\delta, \widehat{Q}_n)e^{4\delta}\|A - B\|\end{aligned}$$

for $A, B \in K_\delta$, and the Lipschitz constant $J(e^\delta, \widehat{Q}_n)e^{4\delta}$ converges to $J(e^\delta, Q)e^{4\delta}$ in probability. This implies that

$$\max_{A\in K_\delta} \big|L_\rho(\exp(A), \widehat{Q}_n) - L_\rho(\exp(A), Q)\big| \to_p 0.$$

In particular,

$$\epsilon_n(\delta) := \min_{A\in\mathbb{W}:\|A\|=\delta} L_\rho(\exp(A), \widehat{Q}_n) \to_p \epsilon(\delta) := \min_{A\in\mathbb{W}:\|A\|=\delta} L(\exp(A), Q) > 0.$$

Whenever $\epsilon_n(\delta) > 0$, we may conclude from Proposition 5.4 the inequality $L_\rho(\exp(A), \widehat{Q}_n) \geq \epsilon_n(\delta)\|A\|/\delta$ for all $A \in \mathbb{W}$ with $\|A\| \geq \delta$. This shows that $L_\rho(\exp(A), \widehat{Q}_n) \to \infty$ as $\|A\| \to \infty$, so $\widehat{Q}_n \in \mathcal{Q}_\rho$ by Proposition 5.5 and Theorem 4.9. Moreover, since $L_\rho(\exp(0), \widehat{Q}_n) = 0$, we may conclude that $\mathbf{\Sigma}_\rho(\widehat{Q}_n) \in \{\exp(A) : A \in K_\delta\}$. □

Proof of Theorem 6.4. According to Theorem 6.3, $\widehat{Q}_n \in \mathcal{Q}_\rho$ with asymptotic probability one. Thus we may replace $\mathcal{L}(\widehat{Q}_n)$ with $\mathcal{L}(\widehat{Q}_n \,|\, \widehat{Q}_n \in \mathcal{Q}_\rho)$ and thus assume that $\widehat{Q}_n \in \mathcal{Q}_n$ almost surely.

As in earlier proofs we define $\mathbb{W} := \mathbb{R}^{q\times q}_{\rm sym}$ in Case 1' and $\mathbb{W} := \{A \in \mathbb{R}^{q\times q}_{\rm sym} : \mathrm{tr}(A) = 0\}$ in Case 0. Since $G_\rho(\widehat{Q}_n) \in \mathbb{W}$, and since $H_\rho(\widehat{Q}_n)$ is a selfadjoint linear operator on the finite-dimensional space $\mathbb{W}$, both $\|G_\rho(\widehat{Q}_n)\|$ and

$$\big\|H_\rho(\widehat{Q}_n) - H_\rho(Q)\big\| \;:=\; \max_{A\in\mathbb{W}\,:\,\|A\|\leq 1} \big\|H_\rho(\widehat{Q}_n)A - H_\rho(Q)A\big\|$$

converge to 0 in probability if, and only if, for arbitrary fixed $A, B \in \mathbb{W}$,

$$\langle A, G_\rho(\widehat{Q}_n)\rangle \;\to_p\; \langle A, G_\rho(Q)\rangle = 0 \quad\text{and}\quad \langle A, H_\rho(\widehat{Q}_n)B\rangle \;\to_p\; \langle A, H_\rho(Q)B\rangle.$$

But this is a consequence of Lemma 8.5: We may write $\langle A, G_\rho(\tilde{Q})\rangle = \int g\, d\tilde{Q}$ and $\langle A, H_\rho(\tilde{Q})B\rangle = \int h\, d\tilde{Q}$ with

$$\begin{aligned} g(M) &:= \mathrm{tr}(A) - \rho'(\mathrm{tr}(M))\,\mathrm{tr}(AM), \\ h(M) &:= \rho'(\mathrm{tr}(M))\,\mathrm{tr}(ABM) + \rho''(\mathrm{tr}(M))\,\mathrm{tr}(AM)\,\mathrm{tr}(BM). \end{aligned}$$

Both $g(M)$ and $h(M)$ are continuous in $M \in \mathbb{Y}$ and satisfy

$$\begin{aligned} |g(M)| &\leq (q + \psi(\mathrm{tr}(M))\|A\|, \\ |h(M)| &\leq \big(\psi(\mathrm{tr}(M)) + \mathrm{tr}(M)^2|\rho''(\mathrm{tr}(M))|\big)\|A\|\|B\| \\ &\leq (2+\kappa)\psi(\mathrm{tr}(M))\|A\|\|B\|, \end{aligned}$$

whence $\int g\, d\widehat{Q}_n \to_p \int g\, dQ$ and $\int h\, d\widehat{Q}_n \to_p \int h\, dQ$.

In particular we may conclude that there exist numbers $\delta_n > 0$ such that $\delta_n \to 0$ and $\mathbb{P}\big(\|G_\rho(\widehat{Q}_n)\| > \delta_n\big) \to 0$. Moreover, with asymptotic probability one, $H_\rho(\widehat{Q}_n)$ is positive definite.

Now we consider $L_\rho(\exp(A), \widehat{Q}_n)$ for $A \in \mathbb{W}$ with $\|A\| \leq \sqrt{\delta_n}$: According to Proposition 5.11,

$$L_\rho(\exp(A), \widehat{Q}_n) \;=\; \langle A, G_\rho(\widehat{Q}_n)\rangle + 2^{-1}H_\rho(A, \widehat{Q}_n) + R_{\rho,2}(A, \widehat{Q}_n).$$

But it follows from Proposition 5.11 that for any fixed $\delta > 0$,

$$\sup_{A\in\mathbb{W}:0<\|A\|\leq\sqrt{\delta_n}} \frac{|R_{\rho,2}(A, \widehat{Q}_n)|}{\|A\|^2} \;\leq\; \Omega(\delta, \widehat{Q}_n)/2 + (\kappa + 1/7)J(\widehat{Q}_n)\delta$$

as soon as $\sqrt{\delta_n} \le \delta$. But $\Omega(\delta, \tilde{Q})/2 + (\kappa + 1/7)J(\tilde{Q}) = \int f_\delta \, d\tilde{Q}$ with

$$f_\delta(M) \;:=\; \sup_{z \in [-\delta, \delta]} \big|\psi_2(e^z \operatorname{tr}(M)) - \psi_2(\operatorname{tr}(M))\big|/2 + (\kappa + 1/7)\psi(\operatorname{tr}(M))\delta.$$

This is continuous in $M \in \mathbb{Y}$, and

$$0 \;\le\; f_\delta(M) \;\le\; (3\kappa/2 + 1/7)\psi(e^\delta \operatorname{tr}(M)) \;\le\; (3\kappa/2 + 1/7)e^{\kappa\delta}\psi(\operatorname{tr}(M)).$$

Hence we may conclude from Lemma 8.5 that

$$\sup_{A \in \mathbb{W}: 0 < \|A\| \le \sqrt{\delta_n}} \frac{|R_{\rho,2}(A, \widehat{Q}_n)|}{\|A\|^2} \;\le\; \int f_\delta \, dQ + o_p(1).$$

But the right hand side converges to 0 as $\delta \to 0$, because $f_\delta(M) \downarrow 0$ as $\delta \downarrow 0$ for any $M \in \mathbb{Y}$. Hence the left hand side converges to 0 in probability.

Together with our considerations about $H_\rho(\widehat{Q}_n)$ we obtain the following expansion:

$$L_\rho(\exp(A), \widehat{Q}_n) \;=\; \langle A, G_\rho(\widehat{Q}_n)\rangle + 2^{-1}\langle A, H_\rho(Q)A\rangle + \widehat{\gamma}_n(A)\|A\|^2$$

where

$$\widehat{\Gamma}_n := \sup_{A \in \mathbb{W}: \|A\| \le \sqrt{\delta_n}} |\widehat{\gamma}_n(A)| \;\to_p\; 0.$$

Now we define

$$\widehat{A}_n \;:=\; -H_\rho(Q)^{-1} G_\rho(\widehat{Q}_n)$$

and note that $c(Q)\|G_\rho(\widehat{Q}_n)\| \le \|\widehat{A}_n\| \le C(Q)\|G_\rho(\widehat{Q}_n)\|$ for suitable constants $0 < c(Q) < C(Q)$. If $\widehat{A}_n = 0$, then $\mathbf{\Sigma}_\rho(\widehat{Q}_n) = I_q$, i.e. $\log(\mathbf{\Sigma}_\rho(\widehat{Q}_n)) = 0$. Thus we focus on the event $\widehat{A}_n \neq 0$. We fix an arbitrary number $\epsilon \in (0, 1)$. For any matrix $A \in \mathbb{W}$ with $\|A - \widehat{A}_n\| = \epsilon\|\widehat{A}_n\|$,

$$\begin{aligned} &L_\rho(\exp(A), \widehat{Q}_n) - L(\exp(\widehat{A}_n), \widehat{Q}_n) \\ &\quad= 2^{-1}\big\langle A - \widehat{A}_n, H_\rho(Q)(A - \widehat{A}_n)\big\rangle + \widehat{\gamma}_n(A)\|A\|^2 - \widehat{\gamma}_n(\widehat{A}_n)\|\widehat{A}_n\|^2. \end{aligned}$$

Note that $\|A\| \le 2\|\widehat{A}_n\|$, and $2\|\widehat{A}_n\| \le \sqrt{\delta_n}$ with asymptotic probability one. In case of $2\|\widehat{A}_n\| \le \sqrt{\delta_n}$,

$$\begin{aligned} \inf_{A \in \mathbb{W}: \|A - \widehat{A}_n\| = \epsilon\|\widehat{A}_n\|} &\big(L_\rho(\exp(A), \widehat{Q}_n) - L(\exp(\widehat{A}_n), \widehat{Q}_n)\big) \\ &\ge \big(2^{-1}\lambda_{\min}(H_\rho(Q))\epsilon^2 - 5\widehat{\Gamma}_n\big)\|\widehat{A}_n\|^2 \\ &= \big(2^{-1}\lambda_{\min}(H_\rho(Q))\epsilon^2 + o_p(1)\big)\|\widehat{A}_n\|^2. \end{aligned}$$

Whenever the right hand side is strictly positive, we may conclude that

$$\big\|\log(\mathbf{\Sigma}_\rho(\widehat{Q}_n)) - \widehat{A}_n\big\| \;\le\; \epsilon\|\widehat{A}_n\| \;\le\; \epsilon C(Q)\|G_\rho(\widehat{Q}_n)\|.$$

These considerations show that $\big\|\log(\mathbf{\Sigma}_\rho(\widehat{Q}_n)) - \widehat{A}_n\big\| \le \epsilon C(Q)\|G_\rho(\widehat{Q}_n)\|$ with asymptotic probability one. Since $\epsilon > 0$ is arbitrarily small, this proves that $\log(\mathbf{\Sigma}_\rho(\widehat{Q}_n))$ equals $\widehat{A}_n + o_p\big(\|G_\rho(\widehat{Q}_n)\|\big)$. □

The proof of Lemma 6.9 relies on the following two propositions involving the Haar distribution on the set of orthogonal matrices in $\mathbb{R}^{q\times q}$. A good reference for Haar distributions in general is the monograph by Eaton (1989).

Proposition 8.6. *Let $U \in \mathbb{R}^{q\times q}$ be a random orthogonal matrix with Haar distribution, i.e. $\mathcal{L}(U) = \mathcal{L}(U^\top) = \mathcal{L}(VU)$ for any fixed orthogonal matrix V. Then for arbitrary indices $i, j, k, \ell, k', \ell' \in \{1, 2, \ldots, q\}$,*

$$\mathbb{E}(U_{ij}^2 U_{k\ell} U_{k'\ell'}) = 0 \quad \text{if } (k,\ell) \neq (k',\ell'), \tag{8.7}$$

$$\mathbb{E}(U_{ij}^4) = c_{q,0} := \frac{3}{q(q+2)}, \tag{8.8}$$

$$\mathbb{E}(U_{ij}^2 U_{i\ell}^2) = \mathbb{E}(U_{ji}^2 U_{\ell i}^2) = c_{q,1} := \frac{1}{q(q+2)} \quad \text{if } j \neq \ell, \tag{8.9}$$

$$\mathbb{E}(U_{ij}^2 U_{k\ell}^2) = c_{q,2} := \frac{q+1}{(q-1)q(q+2)} \quad \text{if } i \neq k, j \neq \ell. \tag{8.10}$$

Proposition 8.7. *Let $M = U\,\mathrm{diag}(\lambda)U^\top$ with a fixed vector $\lambda \in [0,\infty)^q$ and a random orthogonal matrix U as in Proposition 8.6. Then for any matrix $A = A_0 + A_1$ with $A_0 \in \mathbb{W}_0, A_1 \in \mathbb{W}_1$,*

$$\mathbb{E}(\mathrm{tr}(AM)M) = c_0(\lambda)A_0 + c_1(\lambda)A_1,$$

where

$$c_0(\lambda) = \frac{2}{q(q+2)}\Big(\|\lambda\|^2 - \frac{\lambda_+^2 - \|\lambda\|^2}{q-1}\Big) \quad \text{and} \quad c_1(\lambda) = \frac{\lambda_+^2}{q}$$

and $\lambda_+ := \sum_{i=1}^q \lambda_i$.

Proof of Proposition 8.6. By assumption, U has the same distribution as the random matrix $\tilde{U} = (\xi_i\zeta_j U_{ij})_{i,j=1}^q$, where U, ξ and ζ are independent with distribution $\xi, \zeta \sim \mathrm{Unif}(\{-1,1\}^q)$. Hence $U_{ij}^2 U_{k\ell}U_{k'\ell'}$ has the same distribution as the random product $U_{ij}^2 U_{k\ell}U_{k'\ell'}\xi_k\xi_{k'}\zeta_\ell\zeta_{\ell'}$. In case of $(k,\ell) \neq (k',\ell')$, the factor $\xi_k\xi_{k'}\zeta_\ell\zeta_{\ell'}$ is a random sign, and this implies (8.7).

As to the remaining equations, note that U has the same distribution as $U^\top$ and as $\tilde{U} = (U_{\pi(i)\sigma(j)})_{i,j=1}^q$ for arbitrary permutations π, σ of $\{1, 2, \ldots, q\}$. Hence it suffices to show that

$$\mathbb{E}(U_{11}^4) = \frac{3}{q(q+2)}, \tag{8.11}$$

$$\mathbb{E}(U_{11}^2 U_{12}^2) = \frac{1}{q(q+2)}, \tag{8.12}$$

$$\mathbb{E}(U_{11}^2 U_{22}^2) = \frac{q+1}{(q-1)q(q+2)}. \tag{8.13}$$

Any row or column of U is uniformly distributed on the unit sphere of $\mathbb{R}^q$, and this implies that $U_{11}^2 \sim \mathrm{Beta}(a,b)$ with $a = 1/2$, $b = (q-1)/2$. Hence (8.11) follows from

$$\mathbb{E}(U_{11}^4) = \frac{a(a+1)}{(a+b)(a+b+1)} = \frac{3}{q(q+2)}.$$

Now we utilize the fact that all rows of U are unit vectors. Hence

$$1 = \mathbb{E}\Big(\Big(\sum_{j=1}^{q} U_{1j}^2\Big)^2\Big) = \sum_{j,\ell=1}^{q} \mathbb{E}(U_{1j}^2 U_{1\ell}^2) = q\mathbb{E}(U_{11}^4) + q(q-1)\mathbb{E}(U_{11}^2 U_{12}^2)$$
$$= \frac{3}{q+2} + q(q-1)\mathbb{E}(U_{11}^2 U_{12}^2),$$

so

$$\mathbb{E}(U_{11}^2 U_{12}^2) = \frac{1 - 3/(q+2)}{q(q-1)} = \frac{1}{q(q+2)},$$

which is (8.12). Similarly we deduce (8.13):

$$1 = \mathbb{E}\Big(\sum_{j=1}^{q} U_{1j}^2 \sum_{\ell=1}^{q} U_{2\ell}^2\Big) = \sum_{j,\ell=1}^{q} \mathbb{E}(U_{1j}^2 U_{2\ell}^2)$$
$$= q\mathbb{E}(U_{11}^2 U_{12}^2) + q(q-1)\mathbb{E}(U_{11}^2 U_{22}^2)$$
$$= \frac{1}{q+2} + q(q-1)\mathbb{E}(U_{11}^2 U_{22}^2),$$

so

$$\mathbb{E}(U_{11}^2 U_{22}^2) = \frac{1 - 1/(q+2)}{q(q-1)} = \frac{q+1}{(q-1)q(q+2)}.$$ □

Proof of Proposition 8.7. Suppose first that $A = \operatorname{diag}(a)$ for some $a \in \mathbb{R}^q$. Denoting the columns of U with $U_1, U_2, \ldots, U_q$, we may write

$$\mathbb{E}(\operatorname{tr}(AM)M) = \sum_{j=1}^{q} \lambda_j \mathbb{E}(U_j^\top A U_j U \operatorname{diag}(\lambda) U^\top)$$
$$= \sum_{i,j=1}^{q} a_i \lambda_j \mathbb{E}(U_{ij}^2 U \operatorname{diag}(\lambda) U^\top)$$
$$= \sum_{i,j,\ell=1}^{q} a_i \lambda_j \lambda_\ell \mathbb{E}\big(U_{ij}^2 (U_{k\ell} U_{k'\ell})_{k,k'=1}^{q}\big).$$

It follows from Proposition 8.6 that

$$\mathbb{E}\big(U_{ij}^2 (U_{k\ell} U_{k'\ell})_{k,k'=1}^{q}\big)$$
$$= \operatorname{diag}\Big(\big(\mathbb{E}(U_{ij}^2 U_{k\ell}^2)_{k=1}^{q}\Big)$$
$$= \operatorname{diag}\Big(\big(1_{[i=k,j=\ell]} c_{q,0} + 1_{[i=k,j\neq\ell]} c_{q,1} + 1_{[i\neq k,j=\ell]} c_{q,1} + 1_{[i\neq k,j\neq\ell]} c_{q,2}\big)_{k=1}^{q}\Big).$$

Consequently,

$$\mathbb{E}(\operatorname{tr}(AM)M) = \operatorname{diag}(\gamma_1, \gamma_2, \ldots, \gamma_q)$$

with γ_k given by

$$
\begin{aligned}
&\sum_{i,j,\ell=1}^{q} a_i\lambda_j\lambda_\ell\big(1_{[i=k,j=\ell]}c_{q,0} + 1_{[i=k,j\neq\ell]}c_{q,1} + 1_{[i\neq k,j=\ell]}c_{q,1} + 1_{[i\neq k,j\neq\ell]}c_{q,2}\big)\\
&\quad= a_k\|\lambda\|^2c_{q,0} + a_k(\lambda_+^2 - \|\lambda\|^2)c_{q,1}\\
&\qquad + (q\bar a - a_k)\|\lambda\|^2c_{q,1} + (q\bar a - a_k)(\lambda_+^2 - \|\lambda\|^2)c_{q,2}\\
&\quad= \big(\|\lambda\|^2(c_{q,0} - c_{q,1}) + (\lambda_+^2 - \|\lambda\|^2)(c_{q,1} - c_{q,2})\big)\cdot a_k\\
&\qquad + \big(\|\lambda\|^2qc_{q,1} + (\lambda_+^2 - \|\lambda\|^2)qc_{q,2}\big)\cdot\bar a\\
&\quad= \big(\|\lambda\|^2(c_{q,0} - c_{q,1}) + (\lambda_+^2 - \|\lambda\|^2)(c_{q,1} - c_{q,2})\big)\cdot(a_k - \bar a)\\
&\qquad + \big(\|\lambda\|^2(c_{q,0} + (q-1)c_{q,1}) + (\lambda_+^2 - \|\lambda\|^2)(c_{q,1} + (q-1)c_{q,2})\big)\cdot\bar a\\
&\quad= \frac{2}{q(q+2)}\Big(\|\lambda\|^2 - \frac{\lambda_+^2 - \|\lambda\|^2}{q-1}\Big)\cdot(a_k - \bar a) + \frac{\lambda_+^2}{q}\cdot\bar a,
\end{aligned}
$$

where $\lambda_+ := \sum_{i=1}^q \lambda_i$ and $\bar a := q^{-1}\sum_{i=1}^q a_i$. Hence

$$\mathbb{E}(\mathrm{tr}(AM)M) \;=\; c_0(\lambda)\,\mathrm{diag}((a_k - \bar a)_{k=1}^q) + c_1(\lambda)\bar a I_q$$

with $c_0(\lambda), c_1(\lambda)$ as stated.

In general let $A = V\,\mathrm{diag}(a)V^\top$ with an orthogonal matrix $V \in \mathbb{R}^{q\times q}$. Then $A_0 = V\,\mathrm{diag}((a_k - \bar a)_{k=1}^q)V^\top$ and $A_1 = \bar a I_q$, so

$$
\begin{aligned}
\mathbb{E}(\mathrm{tr}(AM)M) \;&=\; V\mathbb{E}\big(\mathrm{tr}(\mathrm{diag}(a)V^\top MV)V^\top MV\big)V^\top\\
&=\; V\big(c_0(\lambda)\,\mathrm{diag}((a_k - \bar a)_{k=1}^q) + c_1(\lambda)\bar a I_q\big)V^\top\\
&=\; c_0(\lambda)A_0 + c_1(\lambda)A_1,
\end{aligned}
$$

because $\mathcal{L}(V^\top MV) = \mathcal{L}\big((V^\top U)\,\mathrm{diag}(\lambda)(V^\top U)^\top\big) = \mathcal{L}(M)$. □

Proof of Lemma 6.9. Let $M \sim Q$ and U be independent, where U is a random orthogonal matrix as in Proposition 8.6. If we write $M = V\,\mathrm{diag}(\Lambda)V^T$ with a random orthogonal matrix $V \in \mathbb{R}^{q\times q}$ and a random vector $\Lambda \in [0,\infty)^q$, then

$$\mathcal{L}(M) = \mathcal{L}(UV\,\mathrm{diag}(\Lambda)V^\top U^\top) = \mathcal{L}\big((UV)\,\mathrm{diag}(\Lambda)(UV)^\top\big) = \mathcal{L}(U\,\mathrm{diag}(\Lambda)U^\top),$$

where the first step follows from orthogonal invariance of Q and the last step follows after conditioning on (Λ, V) and utilizing the fact that $\mathcal{L}(UV) = \mathcal{L}(U)$. Consequently, we may and do assume that $M = U\,\mathrm{diag}(\Lambda)U^\top$. Then, by Proposition 8.7,

$$
\begin{aligned}
H_\rho(Q)A \;&=\; A + \mathbb{E}\big(\rho''(\mathrm{tr}(M))\,\mathrm{tr}(AM)M\big)\\
&=\; A + \mathbb{E}\big(\rho''(\Lambda_+)\,\mathrm{tr}(AM)M\big)\\
&=\; A + \mathbb{E}\big(\rho''(\Lambda_+)\mathbb{E}(\mathrm{tr}(AM)M \mid \Lambda)\big)\\
&=\; A + \mathbb{E}\big(\rho''(\Lambda_+)\big(c_0(\Lambda)A_0 + c_1(\Lambda)A_1\big)\big)\\
&=\; \big(1 + \mathbb{E}\big(\rho''(\Lambda_+)c_0(\Lambda)\big)\big)A_0 + \big(1 + \mathbb{E}\big(\rho''(\Lambda_+)c_1(\Lambda)\big)\big)A_1.
\end{aligned}
$$

Now the assertion follows from the explicit formula for $c_0(\Lambda), c_1(\Lambda)$ and the fact that $\Lambda_+ = \mathrm{tr}(M)$ and $\|\Lambda\|^2 = \|M\|_F^2$. □

Proof of Theorem 6.10. Note first that the nonrandom distributions Q_n satisfy the conditions of Theorem 6.1: It follows from $P_n \to_w P$ that $P_n^{\otimes k} = \mathcal{L}(X_{n1}, \ldots, X_{nk})$ converges weakly to $P^{\otimes k} = \mathcal{L}(X_1, \ldots, X_k)$, where $X_1, \ldots, X_k$ are independent with distribution P. Since the mappings $\mathbb{R}^q \ni x \mapsto xx^\top \in \mathbb{R}^{q\times q}_{\mathrm{sym},\geq 0}$ and $(\mathbb{R}^q)^\ell \ni (x_1, \ldots, x_\ell) \mapsto S(x_1, \ldots, x_\ell)$, $\ell \geq 2$, are continuous, $Q_n \to_w Q$ by the Continuous Mapping Theorem. As to Condition (6.1), note first that for $x \in \mathbb{R}^q$,

$$\psi(\lambda_o \operatorname{tr}(xx^\top)) \;\leq\; \lambda_o^\kappa \psi(\|x\|^2)$$

and for $\ell \geq 2$ points $x_1, \ldots, x_\ell \in \mathbb{R}^q$,

$$\psi\big(\lambda_o \operatorname{tr}(S(x_1, \ldots, x_\ell))\big) \;\leq\; \lambda_o^\kappa (1 - 1/\ell)^{-\kappa} \sum_{i=1}^{\ell} \psi(\|x_\ell\|^2),$$

see also the derivation of (4.9) and Lemma 5.10. Hence we may apply Lemma 8.5 with the non-random triple $\big((\mathbb{R}^q)^k, P_n^{\otimes k}, P^{\otimes k}\big)$ in place of $(\mathbb{Y}, \widehat{Q}_n, Q)$ and the function $\phi(x_1, \ldots, x_k) := \sum_{i=1}^k \psi(\|x_i\|^2)$ to show that under our additional assumptions with $m = 1$,

$$\int \psi(\lambda_o \operatorname{tr}(M))\, Q_n(dM) \;\to\; \int \psi(\lambda_o \operatorname{tr}(M))\, Q(dM).$$

Now we show that the random distributions $\widehat{Q}_n$ satisfy Conditions (6.2) and (6.3) in Theorem 6.3. Because of the preceding considerations for $(Q_n)_n$, it suffices to show that

$$\mathbb{E}\Big|\int g\, d(\widehat{Q}_n - Q_n)\Big| \;\to\; 0 \tag{8.14}$$

whenever $g : \mathbb{Y} \to \mathbb{R}$ is a bounded measurable function or $g(M) = \phi(M) := \psi(\lambda_o \operatorname{tr}(M))$.

In both cases the expected value of $\int g\, d\widehat{Q}_n$ equals $\int g\, dQ_n \in \mathbb{R}$. Consequently, if g is bounded, then

$$\mathbb{E}\Big|\int g\, d(\widehat{Q}_n - Q_n)\Big| \;\leq\; \Big(\operatorname{Var}\Big(\int g\, d\widehat{Q}_n\Big)\Big)^{1/2} \;\leq\; \|g\|_\infty / \sqrt{n/k}.$$

In case of $k = 1$, the latter inequality follows from the well-known identity

$$\operatorname{Var}\Big(\int g\, d\widehat{Q}_n\Big) \;=\; \operatorname{Var}(g(X_{n1}X_{n1}^\top))/n \;\leq\; \|g\|_\infty^2/n.$$

For $k \geq 2$ it follows from inequalities by Hoeffding (1948) for U-statistics, see also Dudley (2002, Section 11.9). This proves (8.14) for bounded g.

In case of $g = \phi$ we fix an arbitrary $R > 0$ and write

$$\mathbb{E}\Big|\int \phi\, d(\widehat{Q}_n - Q_n)\Big| \;\leq\; 2\int (\phi - R)^+\, dQ_n + \mathbb{E}\Big|\int \min\{\phi, R\}\, d(\widehat{Q}_n - Q_n)\Big|$$

$$\begin{aligned} &\le\ 2\int (\phi - R)^+\, dQ_n + R/\sqrt{n/k} \\ &\to\ 2\int (\phi - R)^+\, dQ, \end{aligned}$$

because $(\phi - R)^+ = \phi - \min\{\phi, R\}$. This implies Condition (6.3), because the limit $\int (\phi - R)^+\, dQ$ tends to 0 as $R \to \infty$. □

Proof of Theorem 6.11. As in the proof of Theorem 6.10 it can be shown that

$$\int \psi(\mathrm{tr}(M))^\ell\, Q_n(dM)\ \to\ \int \psi(\mathrm{tr}(M))^\ell\, Q(dM) \quad \text{for } \ell = 1, 2,$$

and that the random distributions $\widehat{Q}_n$ satisfy Conditions (6.2) and (6.4). Hence Theorem 6.4 implies that $\widehat{Q}_n \in \mathcal{Q}_\rho$ with asymptotic probability one, and

$$\sqrt{n}\log(\boldsymbol{\Sigma}_\rho(\widehat{Q}_n))\ =\ H_\rho(Q)^{-1}\big(-\sqrt{n}G_\rho(\widehat{Q}_n)\big) + o_p\big(\sqrt{n}\|G_\rho(\widehat{Q}_n)\|\big).$$

Thus we have to analyze the random matrix

$$\tilde{W}_n\ :=\ -\sqrt{n}G_\rho(\widehat{Q}_n)\ =\ \sqrt{n}\int \big(\rho'(\mathrm{tr}(M))M - I_q\big)\,\widehat{Q}_n(dM)\ \in\ \mathbb{W}$$

in more detail.

In case of $k = 1$ the random matrix $\tilde{W}_n$ equals

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{Z}(X_{ni}) \quad \text{with} \quad \tilde{Z}(x)\ :=\ \rho'(\|x\|^2)xx^\top - I_q$$

for $x \in \mathbb{X}$. Here $\mathbb{E}\tilde{Z}(X_{n1}) = G_\rho(Q_n) = 0$ and $\|\tilde{Z}(\cdot)\|_F \le \psi(\|x\|^2) + \sqrt{q}$. This implies that $\tilde{W}_n = O_p(1)$. Moreover, continuity of ρ' on $(0,\infty)$ and of ψ on $[0,\infty)$ in Case 1' with $\psi(0) = 0$ implies that $\tilde{Z} : \mathbb{X} \to \mathbb{R}^{q\times q}_{\mathrm{sym}}$ is continuous.

In case of $k \ge 2$ we may write

$$\tilde{W}_n\ =\ \sqrt{n}\binom{n}{k}^{-1} \sum_{1\le i_1<\cdots<i_k\le n} M(X_{ni_1}, \ldots, X_{ni_k})$$

with

$$M(x_1, \ldots, x_k)\ :=\ \rho'\big(\mathrm{tr}(S(x_1, \ldots, x_k))\big)S(x_1, \ldots, x_k) - I_q.$$

In Case 0, we define $M(x_1, \ldots, x_k) := 0$ whenever $S(x_1, \ldots, x_k) = 0$. Here

$$\begin{aligned} \|M(x_1, \ldots, x_k)\|_F\ &\le\ \psi\big(\mathrm{tr}(S(x_1, \ldots, x_k))\big) + \sqrt{q} \\ &\le\ (k/(k-1))^\kappa \sum_{i=1}^k \|x_i\|^2 + \sqrt{q}, \end{aligned} \tag{8.15}$$

and $\mathbb{E}M(X_{n1},\dots,X_{nk}) = G_\rho(Q_n) = 0$. Hence standard considerations for U-statistics as in Dudley (2002, Section 11.9), with straightforward extensions to vector- or matrix-valued ones, imply that

$$\tilde{W}_n \;=\; \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{Z}_n(X_{ni}) + o_p(1) \;=\; O_p(1),$$

where

$$\tilde{Z}_n(x) \;:=\; k\mathbb{E}M(x,X_{n2},\dots,X_{nk}) \;=\; k\mathbb{E}(M(X_{n1},X_{n2},\dots,X_{nk})|X_{n1}=x)$$

satisfies $\mathbb{E}\tilde{Z}_n(X_{n1}) = 0$. In addition we define

$$\tilde{Z}(x) \;:=\; k\mathbb{E}M(x,X_2,\dots,X_k).$$

We may conclude from (8.15), continuity of ρ' on $(0,\infty)$ and of ψ on $[0,\infty)$ in Case 1' and dominated convergence that both functions $\tilde{Z}_n$ and $\tilde{Z}$ are continuous on $\mathbb{R}^q$. Further there exists a constant C such that

$$\|\tilde{Z}_n(x)\|_F, \|\tilde{Z}(x)\|_F \;\le\; C + C\psi(\|x\|^2)$$

for all $n \ge k$ and $x \in \mathbb{X}$. Thus it suffices show that

$$\begin{aligned}\mathbb{E}\Big(\Big\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{Z}_n(X_{ni}) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(\tilde{Z}(X_{ni}) - \mathbb{E}\tilde{Z}(X_{n1})\big)\Big\|_F^2\Big)\\ \le\; \mathbb{E}\big(\big\|\tilde{Z}_n(X_{n1}) - \tilde{Z}(X_{n1})\big\|_F^2\big) \;\to\; 0.\end{aligned}$$

To this end we use a well-known result about weak convergence and almost surely convergent representations (Skorohod, 1956; Dudley, 1968): There exists a probability space $(\Omega_o,\mathcal{A}_o,\mathbb{P}_o)$ with random variables $Y \sim P$ and $Y_n \sim P_n$ for $n \ge k$ such that $Y_n \to Y$ almost surely. Now we define $(\Omega,\mathcal{A},\mathbb{P}) := (\Omega_o^k,\mathcal{A}_o^{\otimes k},\mathbb{P}_o^{\otimes k})$ and $X_i(\omega) = Y(\omega_i)$, $X_{ni}(\omega) := Y_n(\omega_i)$ for $1 \le i \le k$, $n \ge k$ and $\omega = (\omega_i)_{i=1}^k \in \Omega$. This construction implies that $(X_{ni})_{i=1}^k \to (X_i)_{i=1}^k$ almost surely. With $\mathcal{A}_*$ denoting the σ-field generated by X_1 and $(X_{n1})_{n\ge k}$ we may write

$$\tilde{Z}_n(X_{n1}) - \tilde{Z}(X_{n1}) \;=\; \mathbb{E}(\tilde{V}_n \,|\, \mathcal{A}_*)$$

with

$$\tilde{V}_n \;:=\; M(X_{n1},X_{n2},\dots,X_{nk}) - M(X_{n1},X_2,\dots,X_k),$$

and

$$\mathbb{E}\big(\big\|\tilde{Z}_n(X_{n1}) - \tilde{Z}(X_{n1})\big\|_F^2\big) \;=\; \mathbb{E}\big(\big\|\mathbb{E}(\tilde{V}_n \,|\, \mathcal{A}_*)\big\|_F^2\big) \;\le\; \mathbb{E}\big(\|\tilde{V}_n\|_F^2\big).$$

But $\tilde{V}_n \to 0$ almost surely, and

$$\|\tilde{V}_n\|_F^2 \;\le\; B_n := C'\sum_{i=1}^{k}\big(\psi(\|X_{ni}\|^2)^2 + \psi(\|X_i\|^2)^2\big)$$

for a suitable constant C'. Furthermore, $B_n \to B := 2C' \sum_{i=1}^k \psi(\|X_i\|^2)^2$ almost surely, and $\mathbb{E}(B_n) \to \mathbb{E}(B) < \infty$. Hence for any fixed $R > 0$,

$$\mathbb{E}\big(\|\tilde{V}_n\|_F^2\big) \;\le\; \mathbb{E}\big(\min\{\|\tilde{V}_n\|_F^2, R\}\big) + \mathbb{E}((B_n - R)^+) \;\to\; \mathbb{E}((B-R)^+),$$

and the right hand side tends to 0 as $R \to \infty$. □

For the proof Remark 6.13 we need an elementary fact about symmetric matrices:

Proposition 8.8. *Let $M \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$ and $x \in \mathbb{R}^q$ such that*

$$BMB^\top \;=\; M \quad \textit{for any orthogonal } B \in \mathbb{R}^{q\times q} \textit{ with } Bx = x.$$

Then there exist real numbers γ, β such that

$$M \;=\; \gamma\, xx^\top + \beta\, I_q.$$

Proof of Proposition 8.8. Let $u \in x^\perp$ with $\|u\| = 1$. Then $B := I_q - 2uu^\top$ defines an orthogonal matrix such that $B^\top = B$, $Bx = x$ and $Bu = -u$. Consequently,

$$u^\top Mx \;=\; u^\top BMB^\top x \;=\; (Bu)^\top M(Bx) \;=\; -u^\top Mx.$$

Hence $Mx \perp x^\perp$ which is equivalent to $Mx = \lambda x$ for some $\lambda \in \mathbb{R}$. In particular, $M(x^\perp) \subset x^\perp$.

Next let u and v be unit vectors in $x^\perp$ such that $u^\top v = 0$ and $Mu = \beta_u u$, $Mv = \beta_v v$ for real numbers β_u, β_v. Then $B := I_q - uu^\top - vv^\top + uv^\top + vu^\top$ defines an orthogonal matrix B such that $B^\top = B$, $Bx = x$, $Bu = v$ and $Bv = u$. Consequently,

$$\beta_u \;=\; u^\top Mu \;=\; (Bu)^\top M(Bu) \;=\; v^\top Mv \;=\; \beta_v.$$

Consequently, there exists a real number β such that $My = \beta y$ for all $y \in x^\perp$.

All in all we obtain the representation $M = \gamma\, xx^\top + \beta\, I_q$, where $\gamma = \lambda\|x\|^{-2} - \beta$ in case of $x \neq 0$. □

Proof of Remark 6.13. Spherical symmetry of P implies that Q is orthogonally invariant. Hence Lemma 6.9 applies to $H_\rho(Q)$, and it suffices to show that $\tilde{Z}(x) = \gamma(\|x\|^2)xx^\top + \beta(\|x\|^2)I_q$ with certain real numbers $\gamma(\|x\|^2)$ and $\beta(\|x\|^2)$. But this is a consequence of Proposition 8.8: For any orthogonal matrix $B \in \mathbb{R}^{q\times q}$,

$$S(Bx, BX_2, \ldots, BX_k) \;=\; BS(x, X_2, \ldots, X_k)B^\top,$$

so it follows from $\mathcal{L}(BX_j) = \mathcal{L}(X_j)$ for $2 \le j \le k$ that

$$\tilde{Z}(Bx) = \mathbb{E}M(Bx, BX_2, \ldots, BX_k) = B\mathbb{E}M(x, X_2, \ldots, X_k)B^\top = B\tilde{Z}(x)B^\top.$$

Restricting our attention temporarily to matrices B such that $Bx = x$ reveals that

$$\tilde{Z}(x) \;=\; \tilde{\gamma}(x)xx^\top + \tilde{\beta}(x)I_q$$

with certain numbers $\tilde{\gamma}(x)$ and $\tilde{\beta}(x)$. But for arbitrary orthogonal $B \in \mathbb{R}^{q\times q}$,

$$\tilde{Z}(Bx) \;=\; \begin{cases} \tilde{\gamma}(Bx)(Bx)(Bx)^\top + \tilde{\beta}(Bx)I_q \;=\; B\big(\tilde{\gamma}(Bx)xx^\top + \tilde{\beta}(Bx)I_q\big)B^\top, \\ B\tilde{Z}(x)B^\top \;=\; B\big(\tilde{\gamma}(x)xx^\top + \tilde{\beta}(x)I_q\big)B^\top, \end{cases}$$

whence

$$\tilde{\gamma}(Bx)xx^\top + \tilde{\beta}(Bx)I_q \;=\; \tilde{\gamma}(x)xx^\top + \tilde{\beta}(x)I_q.$$

Multiplying the latter equation with $y^\top$ from the left and with y from the right, where $0 \neq y \in x^\perp$, reveals that $\tilde{\beta}(Bx) = \tilde{\beta}(x)$, i.e. $\tilde{\beta}(x) = \beta(\|x\|^2)$. Then multiplication with $x^\top$ from the left and x from the right reveals that $\tilde{\gamma}(Bx) = \tilde{\gamma}(x)$, i.e. $\tilde{\gamma}(x) = \gamma(\|x\|^2)$. □

Proof of Remark 6.14. If P is spherically symmetric around 0, we may represent a random vector $X \sim P$ as $X = RU$ with independent random variables $R \geq 0$ and $U \in \mathbb{R}^q$, where U is uniformly distributed on the unit sphere of $\mathbb{R}^q$.

In case of $\nu = 0$ (Case 0) we know already that

$$H_\rho(Q)A \;=\; \frac{q}{q+2}\,A_0$$

for any $A = A_0 + aI_q \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$, where $a = q^{-1}\operatorname{tr}(A)$ and $\operatorname{tr}(A_0) = 0$. Hence

$$\begin{aligned} Z(x) \;&=\; H_\rho(Q)^{-1}\big(q\|x\|^{-2}xx^\top - I_q\big) \;=\; \frac{q}{\|x\|^2}H_\rho(Q)^{-1}A_0(x) \;=\; \frac{q+2}{\|x\|^2}A_0(x) \\ &=\; (\nu + \|x\|^2)^{-1}\big(c_0A_0(x) + c_1a(x)I_q\big) \end{aligned}$$

with $\nu = 0$ and c_0, c_1 as stated. Note that $c_1 = 0$ when $\nu = 0$.

In case of $\nu > 0$ (Case 1'), Proposition 8.7, applied with $\lambda = (1, 0, \ldots, 0)^\top$, entails that $A = A_0 + aI_q$ as above is mapped to

$$\begin{aligned} H_\rho(Q)A \;&=\; A - \mathbb{E}\Big(\frac{(\nu+q)R^4}{(\nu+R^2)^2}\,U^\top AU\,UU^\top\Big) \\ &=\; A - \mathbb{E}\Big(\frac{(\nu+q)R^4}{(\nu+R^2)^2}\Big)\Big(\frac{2}{q(q+2)}\,A_0 + \frac{1}{q}\,aI_q\Big) \\ &=\; \Big(1 - \frac{2(q-\nu+\beta\nu)}{q(q+2)}\Big)\,A_0 + \Big(1 - \frac{q-\nu+\beta\nu}{q}\Big)aI_q \\ &=\; \frac{q + 2\nu(1-\beta)/q}{q+2}\,A_0 + \frac{\nu(1-\beta)}{q}\,aI_q, \end{aligned}$$

because

$$\mathbb{E}\Big(\frac{(\nu+q)R^4}{(\nu+R^2)^2}\Big) \;=\; \mathbb{E}\Big(\frac{(\nu+q)(R^2-\nu)}{\nu+R^2} + \frac{(\nu+q)\nu^2}{(\nu+R^2)^2}\Big) \;=\; q - (1-\beta)\nu$$

by the definition of β and since $\Sigma_\rho(Q) = I_q$. Note that the latter implies the equations $\mathbb{E}(R^2/(\nu+R^2)) = q/(\nu+q)$ and $\mathbb{E}(1/(\nu+R^2)) = 1/(\nu+q)$. Consequently,

$$H_\rho(Q)^{-1}A \;=\; \frac{q+2}{q+2(1-\beta)\nu/q}\,A_0 + \frac{q}{(1-\beta)\nu}\,aI_q.$$

This yields the representation

$$\begin{aligned}
Z(x) \;&=\; H_\rho(Q)^{-1}\big(\rho'(\|x\|^2)xx^\top - I_q\big)\\
&=\; (\nu+\|x\|^2)^{-1}H_\rho(Q)^{-1}\big((\nu+q)(A_0(x)+a(x)I_q+I_q) - (\nu+\|x\|^2)I_q\big)\\
&=\; (\nu+\|x\|^2)^{-1}H_\rho(Q)^{-1}\big((\nu+q)A_0(x)+\nu a(x)I_q\big)\\
&=\; (\nu+\|x\|^2)^{-1}\Big(\frac{(\nu+q)(q+2)}{q+2(1-\beta)\nu/q}\,A_0(x) + \frac{q}{1-\beta}\,a(x)I_q\Big)\\
&=\; (\nu+\|x\|^2)^{-1}\big(c_0A_0(x)+c_1a(x)I_q\big)
\end{aligned}$$

with c_0 and c_1 as stated. □

8.5. Proofs for Section 7

Proof of Theorem 7.1. Note that $\tilde L(\cdot,\tilde P)$ is equal to the scatter-only functional $L(\cdot,Q)$ with $Q = Q^1(\tilde P) = \mathcal{L}\big(y(X)y(X)^\top\big)$, $X\sim P$. In what follows let $\mathbb{H}_0 := \big\{(x^\top,0)^\top : x\in\mathbb{R}^q\big\}$ and $\mathbb{H}_1 := \big\{(x^\top,1)^\top : x\in\mathbb{R}^q\big\} = \{y(x) : x\in\mathbb{R}^q\}$. For any linear subspace $\mathbb{W}$ of $\mathbb{R}^{q+1}$ with $1\le\dim(\mathbb{W})\le q$, elementary linear algebra reveals that either $\mathbb{W}\subset\mathbb{H}_0$ or

$$\mathbb{W}\cap\mathbb{H}_1 \;=\; \big\{y(a+v) : v\in\mathbb{V}\big\}$$

for some $a\in\mathbb{R}^q$ and a linear subspace $\mathbb{V}$ of $\mathbb{R}^q$ with $\dim(\mathbb{V}) = \dim(\mathbb{W})-1$.

In case of $\nu = 1$, we know from Theorem 4.9 that $\tilde L(\cdot,\tilde P) = L(\cdot,Q)$ possesses a unique minimizer up to multiplication with positive scalars if, and only if,

$$Q(\mathbb{M}(\mathbb{W})) \;=\; \tilde P(\mathbb{W}) \;<\; \frac{\dim(\mathbb{W})}{q+1}$$

for arbitrary linear subspaces $\mathbb{W}$ of $\mathbb{R}^{q+1}$ with $1\le\dim(\mathbb{W})\le q$. In view of the previous considerations, and since $\tilde P(\mathbb{H}_0) = 0$, this is equivalent to

$$P(a+\mathbb{V}) \;<\; \frac{\dim(\mathbb{V})+1}{q+1}$$

for arbitrary $a\in\mathbb{R}^q$ and any linear subspace $\mathbb{V}$ of $\mathbb{R}^q$ with $0\le\dim(\mathbb{V})<q$.

In case of $\nu>1$, we apply Theorem 4.9 to $\rho(s) = \rho_{\nu-1,q+1}(s) = (\nu+q)\log(\nu+s-1)$, i.e. $\psi(\infty) = \nu+q$. Hence $\tilde L(\cdot,\tilde P) = L(\cdot,Q)$ possesses a unique minimizer $\Gamma\in\mathbb{R}^{(q+1)\times(q+1)}_{\mathrm{sym},>0}$ if, and only if

$$Q(\mathbb{M}(\mathbb{W})) \;=\; \tilde P(\mathbb{W}) \;<\; \frac{\dim(\mathbb{W})+\nu-1}{q+\nu}$$

for arbitrary linear subspaces $\mathbb{W}$ of $\mathbb{R}^{q+1}$ with $0 \le \dim(\mathbb{W}) \le q$. Since $\tilde{P}(\{0\}) = 0$, it suffices to consider the case $\dim(\mathbb{W}) \ge 1$, and then the previous considerations show that our requirement on $\tilde{P}$ is equivalent to

$$P(a + \mathbb{V}) \;<\; \frac{\dim(\mathbb{V}) + \nu}{q + \nu}$$

for arbitrary $a \in \mathbb{R}^q$ and any linear subspace $\mathbb{V}$ of $\mathbb{R}^q$ with $0 \le \dim(\mathbb{V}) < q$.

It remains to show that for $\nu > 1$, a minimizer Γ of $\tilde{L}(\cdot, \tilde{P})$ satisfies $\Gamma_{q+1,q+1} = 1$. To this end, recall that Γ satisfies the fixed-point equation

$$\Gamma \;=\; \Psi(Q) = \mathbb{E}\Big(\frac{q+\nu}{Y^\top \Gamma^{-1} Y + \nu - 1}\, YY^\top\Big) \tag{8.16}$$

with $Y := y(X)$, $X \sim P$. In particular, since $Y_{q+1} = 1$ almost surely,

$$\Gamma_{q+1,q+1} \;=\; \mathbb{E}\frac{q+\nu}{Y^\top \Gamma^{-1} Y + \nu - 1}.$$

But (8.16) implies also that

$$\begin{aligned} q+1 \;=\; \operatorname{tr}(\Gamma^{-1}\Psi(Q)) \;&=\; \mathbb{E}\frac{(q+\nu) Y^\top \Gamma^{-1} Y}{Y^\top \Gamma^{-1} Y + \nu - 1} \\ &=\; q + \nu - \mathbb{E}\frac{(q+\nu)(\nu-1)}{Y^\top \Gamma^{-1} Y + \nu - 1} \\ &=\; q + \nu - (\nu-1)\Gamma_{q+1,q+1} \\ &=\; q + 1 + (\nu-1)(1 - \Gamma_{q+1,q+1}), \end{aligned}$$

i.e. $\Gamma_{q+1,q+1} = 1$. □

Proof of Theorem 7.2. It follows from Theorem 6.4 that with asymptotic probability one there exists a unique minimizer $\mathbf{\Gamma}(\widehat{P}_n)$ of

$$\int \big[\rho(y(x)^\top \Gamma^{-1} y(x)) - \rho(\|y(x)\|^2)\big]\, \widehat{P}_n(dx) + \log\det(\Gamma)$$

over all $\Gamma \in \mathbb{R}^{(q+1)\times(q+1)}_{\mathrm{sym},>0}$. In case of $\nu = 1$ we also require that $\det(\Gamma) = 1$. Moreover,

$$\mathbf{\Gamma}(\widehat{P}_n) \;=\; I_{q+1} - \tilde{H}(P)^{-1}\tilde{G}(\widehat{P}_n) + o_p\big(\|\tilde{G}(\widehat{P}_n)\|\big)$$

with the operator $\tilde{H}(P)$ as stated, and

$$\tilde{G}(\widehat{P}_n) \;:=\; I_{q+1} - \int \tilde{\rho}'(\|y(x)\|^2) y(x) y(x)^\top\, \widehat{P}_n(dx) \;\to_p\; 0.$$

Now we set

$$\tilde{Z}(x) \;:=\; \tilde{H}(P)^{-1}\big(\rho'(\|x\|^2) y(x) y(x)^\top - I_{q+1}\big)$$

for $x \in \mathbb{R}^q$. This defines a bounded, continuous function $\tilde{Z} : \mathbb{R}^q \to \mathbb{R}^{(q+1)\times(q+1)}_{\mathrm{sym}}$ with $\int \tilde{Z}\, dP = 0$. Since the operator $\tilde{H}(P)$ is non-singular, both $\|\tilde{G}(\widehat{P}_n)\|$ and

$\delta_n := \bigl\|\int \tilde{Z}\, d\widehat{P}_n\bigr\|$ tend to zero in probability at the same speed, and we may write

$$\mathbf{\Gamma}(\widehat{P}_n) \;=\; I_{q+1} + \int \tilde{Z}\, d\widehat{P}_n + o_p(\delta_n).$$

But then

$$\begin{aligned}
&\begin{bmatrix} \mathbf{\Sigma}(\widehat{P}_n) + \boldsymbol{\mu}(\widehat{P}_n)\boldsymbol{\mu}(\widehat{P}_n)^\top & \boldsymbol{\mu}(\widehat{P}_n) \\ \boldsymbol{\mu}(\widehat{P}_n)^\top & 1 \end{bmatrix} \\
&\quad= \; (\mathbf{\Gamma}(\widehat{P}_n)_{q+1,q+1})^{-1}\mathbf{\Gamma}(\widehat{P}_n) \\
&\quad= \; \Bigl(1 - \int \tilde{Z}_{q+1,q+1}\, d\widehat{P}_n + o_p(\delta_n)\Bigr)\Bigl(I_{q+1} + \int \tilde{Z}\, d\widehat{P}_n + o_p(\delta_n)\Bigr) \\
&\quad= \; I_{q+1} + \int \bigl(\tilde{Z} - \tilde{Z}_{q+1,q+1} I_{q+1}\bigr)\, d\widehat{P}_n + o_p(\delta_n).
\end{aligned}$$

In particular, $\boldsymbol{\mu}(\widehat{P}_n) = O_p(\delta_n)$, whence $\boldsymbol{\mu}(\widehat{P}_n)\boldsymbol{\mu}(\widehat{P}_n)^\top = O_p(\delta_n^2) = o_p(\delta_n)$ and thus

$$\begin{bmatrix} \mathbf{\Sigma}(\widehat{P}_n) - I_q & \boldsymbol{\mu}(\widehat{P}_n) \\ \boldsymbol{\mu}(\widehat{P}_n)^\top & 0 \end{bmatrix} \;=\; \int \bigl(\tilde{Z} - \tilde{Z}_{q+1,q+1} I_{q+1}\bigr)\, d\widehat{P}_n + o_p(\delta_n).$$

It remains to show that $\tilde{Z}_{q+1,q+1}(x) = 0$ for any fixed $x \in \mathbb{R}^q$ in case of $\nu > 1$. To this end we consider the nonrandom distributions $P_n := (1-n^{-1})P + n^{-1}\delta_x$. For sufficiently large n, $\mathbf{\Gamma}(P_n)$ is well-defined with $\mathbf{\Gamma}(P_n)_{q+1,q+1} = 1$. On the other hand, $\int \tilde{Z}\, dP_n = n^{-1}\tilde{Z}(x)$ and

$$\mathbf{\Gamma}(P_n) \;=\; I_{q+1} + n^{-1}\tilde{Z}(x) + o(n^{-1}),$$

which implies that $\tilde{Z}_{q+1,q+1}(x) = 0$. □

Proof of Remarks 7.4 and 7.5. Recall that $\tilde{G}(P) = 0$ is equivalent to

$$\int \frac{\nu+q}{\nu+\|x\|^2} \begin{bmatrix} xx^\top & x \\ x^\top & 1 \end{bmatrix} P(dx) \;=\; \begin{bmatrix} I_q & 0 \\ 0 & 1 \end{bmatrix}, \tag{8.17}$$

in particular,

$$\int \frac{(\nu+q)\|x\|^2}{\nu+\|x\|^2}\, P(dx) \;=\; q \quad\text{and}\quad \int \frac{\nu+q}{\nu+\|x\|^2}\, P(dx) \;=\; 1. \tag{8.18}$$

Now we introduce the auxiliary objects

$$\begin{aligned}
\Psi_2 = \Psi_2(P) \;&:=\; \int \frac{\nu+q}{(\nu+\|x\|^2)^2}\, xx^\top\, P(dx) \;\in\; \mathbb{R}^{q\times q}_{\mathrm{sym},>0}, \\
\beta = \beta(P) \;&:=\; \int \frac{(\nu+q)\nu}{(\nu+\|x\|^2)^2}\, P(dx) \;>\; 0,
\end{aligned}$$

i.e.

$$\beta + \operatorname{tr}(\Psi_2) \;=\; \int \frac{\nu+q}{\nu+\|x\|^2} P(dx) \;=\; 1,$$

and the operator $H = H(P) : \mathbb{R}^{q\times q}_{\mathrm{sym}} \to \mathbb{R}^{q\times q}_{\mathrm{sym}}$ given by

$$HA \;:=\; A - \int \frac{\nu+q}{(\nu+\|x\|^2)^2}\, x^\top A x\, xx^\top\, P(dx).$$

Then for a matrix

$$M \;=\; \begin{bmatrix} A & b \\ b^\top & c \end{bmatrix}$$

with $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$, $b \in \mathbb{R}^q$ and $c \in \mathbb{R}$, we may write

$$\begin{aligned} \tilde H(P)M \;&=\; \begin{bmatrix} A & b \\ b^\top & c \end{bmatrix} - \int \frac{(\nu+q)(x^\top A x + 2x^\top b + c)}{(\nu+\|x\|^2)^2} \begin{bmatrix} xx^\top & x \\ x^\top & 1 \end{bmatrix} P(dx) \\ &=\; \begin{bmatrix} HA - c\Psi_2 & 0 \\ 0 & (1-\beta/\nu)c - \langle \Psi_2, A\rangle \end{bmatrix} + \begin{bmatrix} 0 & (I_q - 2\Psi_2)b \\ b^\top(I_q - 2\Psi_2) & 0 \end{bmatrix} \end{aligned}$$

Here we utilized the fact that any term of the form $f(xx^\top)x$ integrates to 0, due to the symmetry of P. Consequently,

$$\begin{gathered} \left\{ \tilde H(P) \begin{bmatrix} A & 0 \\ 0 & c \end{bmatrix} : A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}, c \in \mathbb{R} \right\} \;\subset\; \left\{ \begin{bmatrix} A & 0 \\ 0 & c \end{bmatrix} : A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}, c \in \mathbb{R} \right\} \\ \left\{ \tilde H(P) \begin{bmatrix} 0 & b \\ b^\top & 0 \end{bmatrix} : b \in \mathbb{R}^q \right\} \;=\; \left\{ \begin{bmatrix} 0 & b \\ b^\top & 0 \end{bmatrix} : b \in \mathbb{R}^q \right\}, \end{gathered}$$

where the latter equality follows from $\tilde H(P)$ being nonsingular on $\tilde{\mathbb{M}}$. In particular, $B = B(P) := (I_q - 2\Psi_2(P))^{-1} \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$ exists, and

$$\begin{aligned} \tilde Z(x) \;&=\; \tilde H(P)^{-1}\big(\rho'(\|x\|^2) y(x) y(x)^{-1} - I_{q+1}\big) \\ &=\; \tilde H(P)^{-1} \begin{bmatrix} \rho'(\|x\|^2) xx^\top - I_q & 0 \\ 0 & \rho'(\|x\|^2) - 1 \end{bmatrix} + \rho'(\|x\|^2) \tilde H(P)^{-1} \begin{bmatrix} 0 & x \\ x^\top & 0 \end{bmatrix} \\ &=\; \begin{bmatrix} Z(xx^\top) & 0 \\ 0 & z(\|x\|^2) \end{bmatrix} + \rho'(\|x\|^2) \begin{bmatrix} 0 & Bx \\ x^\top B & 0 \end{bmatrix} \end{aligned}$$

with certain bounded, continuous functions $Z : \mathbb{R}^{q\times q}_{\mathrm{sym},\geq 0} \to \mathbb{R}^{q\times q}_{\mathrm{sym}}$ and $z : [0,\infty) \to \mathbb{R}$.

This proves Remark 7.4. In the special case of P being spherically symmetric around 0, a random vector $X \sim P$ may be written as $X = RU$ with independent random variables $R \geq 0$ and $U \in \mathbb{R}^q$, where U is uniformly distributed on the unit sphere of $\mathbb{R}^q$. Then (8.18) and the definition of β translate to

$$\mathbb{E}\Big(\frac{(\nu+q)R^2}{\nu+R^2}\Big) \;=\; q, \quad \mathbb{E}\Big(\frac{\nu+q}{\nu+R^2}\Big) \;=\; 1 \quad \text{and} \quad \beta \;=\; \mathbb{E}\Big(\frac{(\nu+q)\nu}{(\nu+R^2)^2}\Big).$$

Further, it follows from $\mathbb{E}(UU^\top) = q^{-1} I_q$ that

$$\Psi_2 \;=\; \mathbb{E}\Big(\frac{(\nu+q)R^2}{(\nu+R^2)^2}\, UU^\top\Big) \;=\; \mathbb{E}\Big(\frac{\nu+q}{\nu+R^2} - \frac{(\nu+q)\nu}{(\nu+R^2)^2}\Big)\,\frac{1}{q}\, I_q \;=\; \gamma_1\, I_q$$

with

$$\gamma_1 \;:=\; \frac{1-\beta}{q}.$$

Now we write $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$ as $A = A_0 + aI_q$ with $a := \mathrm{tr}(A)/q$, so $\mathrm{tr}(A_0) = 0$. Then

$$\langle \Psi_2, A\rangle \;=\; \frac{1-\beta}{q}\,\mathrm{tr}(A) \;=\; \gamma_1 q a$$

and, as shown in the proof of Remark 6.14,

$$H(P)A \;=\; \gamma_0\, A_0 + \gamma_1 \nu\, a I_q,$$

with

$$\gamma_0 \;:=\; \frac{q+2\gamma_1\nu}{q+2}.$$

Hence for $A_0 \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$ with $\mathrm{tr}(A_0) = 0$, $a \in \mathbb{R}$, $b \in \mathbb{R}^q$ and $c \in \mathbb{R}$,

$$\tilde H(P)\begin{bmatrix} A_0 + aI_q & b \\ b^\top & c \end{bmatrix} \;=\; \begin{bmatrix} \gamma_0 A_0 + \gamma_1(\nu a - c) I_q & (1-2\gamma_1) b \\ (1-2\gamma_1) b^\top & (1-\beta/\nu)c - \gamma_1 q a \end{bmatrix}.$$

In case of $\nu = 1$ we only consider the case $\mathrm{tr}(M) = 0$, i.e. $c = -qa$. Then

$$\tilde H(P)\begin{bmatrix} A_0 + aI_q & b \\ b^\top & -qa \end{bmatrix} \;=\; \begin{bmatrix} \gamma_0 A_0 + \gamma_1(q+1) a I_q & (1-2\gamma_1) b \\ (1-2\gamma_1) b^\top & -q\gamma_1(q+1)a \end{bmatrix},$$

and this shows that

$$\tilde H(P)^{-1}\begin{bmatrix} A_0 + aI_q & b \\ b^\top & -qa \end{bmatrix} \;=\; \begin{bmatrix} \gamma_0^{-1} A_0 + \gamma_1^{-1}(q+1)^{-1} a I_q & (1-2\gamma_1)^{-1} b \\ (1-2\gamma_1)^{-1} b^\top & -q\gamma_1^{-1}(q+1)^{-1} a \end{bmatrix}.$$

Now we consider $x \in \mathbb{R}^q$ and write $xx^\top = A_0(x) + a(x) I_q + I_q$ with $a(x) := q^{-1}\|x\|^2 - 1$, so $\mathrm{tr}(A_0(x)) = 0$. Then

$$\begin{aligned}
\tilde Z(x) \;&=\; (1+\|x\|^2)^{-1} \tilde H(P)^{-1}\big((1+q) y(x) y(x)^\top - (1+\|x\|^2) I_{q+1}\big) \\
&=\; (1+\|x\|^2)^{-1} \tilde H(P)^{-1} \begin{bmatrix} (1+q)A_0(x) + a(x) I_q & (1+q)x \\ (1+q)x^\top & -qa(x) \end{bmatrix} \\
&=\; (1+\|x\|^2)^{-1} \begin{bmatrix} \dfrac{1+q}{\gamma_0} A_0(x) + \dfrac{1}{\gamma_1(1+q)} a(x) I_q & (1-2\gamma_1)^{-1} x \\ (1-2\gamma_1)^{-1} x^\top & \dfrac{-q}{\gamma_1(1+q)} a(x) \end{bmatrix}.
\end{aligned}$$

Consequently,

$$\tilde Z(x) - \tilde Z(x)_{q+1,q+1} I_{q+1} \;=\; (1+\|x\|^2)^{-1} \begin{bmatrix} c_0 A_0(x) + c_1 a(x) I_q & c_2 x \\ c_2 x^\top & 0 \end{bmatrix}$$

with

$$\begin{aligned} c_0 &:= \frac{1+q}{\gamma_0} = \frac{(q+1)(q+2)}{q+2(1-\beta)/q}, \\ c_1 &:= \frac{1}{\gamma_1} = \frac{q}{1-\beta}, \\ c_2 &:= (1-2\gamma_1)^{-1} = \frac{q}{q-2(1-\beta)}. \end{aligned}$$

In case of $\nu > 1$, elementary calculations reveal that the inverse of the mapping

$$\begin{bmatrix} a \\ c \end{bmatrix} \mapsto \begin{bmatrix} \gamma_1(\nu a - c) \\ (1-\beta/\nu)c - \gamma_1 q a \end{bmatrix} = \begin{bmatrix} \gamma_1\nu & -\gamma_1 \\ -\gamma_1 q & 1-\beta/\nu \end{bmatrix} \begin{bmatrix} a \\ c \end{bmatrix}$$

is given by

$$\begin{bmatrix} a \\ c \end{bmatrix} \mapsto \frac{1}{\nu-1} \begin{bmatrix} (1-\beta/\nu)/\gamma_1 & 1 \\ q & 1\nu \end{bmatrix} \begin{bmatrix} a \\ c \end{bmatrix}.$$

Consequently

$$\tilde{H}(P)^{-1} \begin{bmatrix} A_0 + aI_q & b \\ b^\top & c \end{bmatrix} = \begin{bmatrix} \gamma_0^{-1} A_0 + \dfrac{(1-\beta/\nu)\gamma_1^{-1} a + c}{\nu-1} I_q & (1-2\gamma_1)^{-1} b \\ (1-2\gamma_1)^{-1} b^\top & \dfrac{qa + \nu c}{\nu-1} \end{bmatrix}.$$

Hence

$$\begin{aligned} \tilde{Z}(x) &= (\nu + \|x\|^2)^{-1} \tilde{H}(P)^{-1} \big((\nu+q) y(x) y(x)^\top - (\nu + \|x\|^2) I_{q+1} \big) \\ &= (\nu + \|x\|^2)^{-1} \tilde{H}(P)^{-1} \begin{bmatrix} (\nu+q)A_0(x) + \nu a(x) I_q & (\nu+q)x \\ (\nu+q)x^\top & -qa(x) \end{bmatrix} \\ &= (1 + \|x\|^2)^{-1} \begin{bmatrix} \dfrac{\nu+q}{\gamma_0} A_0(x) + \dfrac{q}{1-\beta} a(x) I_q & (1-2\gamma_1)^{-1} x \\ (1-2\gamma_1)^{-1} x^\top & 0 \end{bmatrix} \\ &= (1 + \|x\|^2)^{-1} \begin{bmatrix} c_0 A_0(x) + c_1 a(x) I_q & c_2 x \\ c_2 x^\top & 0 \end{bmatrix} \end{aligned}$$

with c_0, c_1, c_2 as stated. □

Acknowledgement

Constructive comments by an associate editor and two referees are gratefully acknowledged. Many thanks to David Tyler for stimulating discussions, in particular for encouraging us to drop the assumption of ρ' being non-increasing.

List of notation and assumptions

Linear and affine transformations Let P and Q be probability distributions on $\mathbb{R}^q$ and $\mathbb{R}^{q\times q}_{\mathrm{sym},\geq 0}$, respectively. For $a \in \mathbb{R}^q$, $B \in \mathbb{R}^{q\times q}_{\mathrm{ns}}$ and $X \sim P$, $S \sim Q$,

$$P^B := \mathcal{L}(BX), \quad P^{a,B} := \mathcal{L}(a + BX),$$

and

$$Q^B \;:=\; \mathcal{L}(BSB^\top), \quad Q_B \;:=\; \mathcal{L}(B^{-1}SB^{-\top}).$$

Special (empirical) distributions Let $X = X_1, X_2, X_3, \ldots$ be i.i.d. $\sim P$. Then for $k \geq 2$,

$$Q^1(P) \;:=\; \mathcal{L}(XX^\top) \quad \text{and} \quad Q^k(P) \;:=\; \mathcal{L}\big(S(X_1, X_2, \ldots, X_k)\big)$$

with $S(x_1, x_2, \ldots, x_k)$ denoting the sample covariance matrix of $x_1, x_2, \ldots, x_k \in \mathbb{R}^q$. Furthermore,

$$\widehat{P} \;:=\; n^{-1}\sum_{i=1}^n \delta_{X_i}, \quad \widehat{Q}^1 \;:=\; n^{-1}\sum_{i=1}^n \delta_{X_i X_i^\top}$$

and

$$\widehat{Q}^k \;:=\; \binom{n}{k}^{-1} \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} \delta_{S(X_{i_1}, X_{i_2}, \ldots, X_{i_k})}.$$

Log-likelihood functions (times -2) and derivatives

$$\begin{aligned} L(\mu, \Sigma, P) \;&:=\; \int \big[\rho\big((x-\mu)^\top\Sigma^{-1}(x-\mu)\big) - \rho(x^\top x)\big]\, P(dx) + \log\det(\Sigma), \\ L_\rho(\Sigma, Q) \;&:=\; \int \big[\rho(\mathrm{tr}(\Sigma^{-1}M)) - \rho(\mathrm{tr}(M))\big]\, Q(dM) + \log\det\Sigma. \end{aligned}$$

Under certain conditions, as $\mathbb{R}^{q\times q}_{\mathrm{sym}} \ni A \to 0$,

$$\begin{aligned} L_\rho(\exp(A), Q) \;&=\; \langle G_\rho(Q), A\rangle + o(\|A\|) \\ &=\; \langle G_\rho(Q), A\rangle + 2^{-1}H_\rho(A, Q) + o(\|A\|^2), \end{aligned}$$

where

$$\begin{aligned} G_\rho(Q) \;&:=\; I_q - \Psi_\rho(Q), \quad \Psi_\rho(Q) \;:=\; \Psi_\rho(I_q, Q), \\ \Psi_\rho(\Sigma, Q) \;&:=\; \int \rho'(\mathrm{tr}(\Sigma^{-1}M))M\, Q(dM), \\ H_\rho(A, Q) \;&:=\; \int \big(\rho'(\mathrm{tr}(M))\,\mathrm{tr}(A^2M) + \rho''(\mathrm{tr}(M))\,\mathrm{tr}(AM)^2\big)\, Q(dM). \end{aligned}$$

Moreover, $H_\rho(A, Q) = \langle H_\rho(Q)A, A\rangle$ with the linear operator $H_\rho(Q) : \mathbb{R}^{q\times q}_{\mathrm{sym}} \to \mathbb{R}^{q\times q}_{\mathrm{sym}}$ given by

$$H_\rho(Q)A \;:=\; 2^{-1}\big(A\Psi_\rho(Q) + \Psi_\rho(Q)A\big) + \int \rho''(\mathrm{tr}(M))\,\mathrm{tr}(AM)M\, Q(dM).$$

Sometimes we write $\mathbb{R}^{q\times q}_{\mathrm{sym}} = \mathbb{W}_0 \oplus \mathbb{W}_1$ with

$$\mathbb{W}_0 \;:=\; \{A \in \mathbb{R}^{q\times q}_{\mathrm{sym}} : \mathrm{tr}(A) = 0\} \quad \text{and} \quad \mathbb{W}_1 \;:=\; \{sI_q : s \in \mathbb{R}\}.$$

In Case 0, we view $H_\rho(Q)$ as an endomorphism of $\mathbb{W}_0$.

Assumptions on ρ and Q We assume that ρ is continuously differentiable on $(0,\infty)$ with derivative $\rho' > 0$. For $s > 0$ we define

$$\psi(s) \;:=\; s\rho'(s).$$

Case 0 $\rho(s) = q\log(s)$ for $s > 0$, and $Q(\{0\}) = 0$.

Case 1 ψ is strictly increasing on $(0,\infty)$ with limits $\psi(0) = 0$ and $\psi(\infty) \in (q,\infty]$. Moreover, $J_\rho(\lambda, Q) := \int \psi(\lambda \operatorname{tr}(M))\, Q(dM) < \infty$ for any $\lambda \geq 1$.

Case 1' ρ is twice continuously differentiable on $(0,\infty)$ with $\psi' > 0$, and ψ has limits $\psi(0) = 0$ and $\psi(\infty) \in (q,\infty]$. Moreover, $J_\rho(Q) := \int \psi(\operatorname{tr}(M))\, Q(dM) < \infty$, and there exists a constant $\kappa \geq 0$ such that $s\psi'(s) \leq \kappa\psi(s)$ for all $s > 0$.

Existence of $\boldsymbol{\Sigma}_\rho(Q)$ Let $\mathcal{Q}_\rho$ be the set of all distributions Q such that $L_\rho(\cdot, Q)$ is real-valued and has a unique minimizer $\boldsymbol{\Sigma}_\rho(Q) \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$, where $\det(\boldsymbol{\Sigma}_\rho(Q)) = 1$ in Case 0. To characterize $\mathcal{Q}_\rho$ let

$$\begin{aligned} \mathcal{V}_q \;&:=\; \{\mathbb{V} : \mathbb{V} \text{ a linear subspace of } \mathbb{R}^q\}, \\ \mathbb{M}(\mathbb{V}) \;&:=\; \left\{M \in \mathbb{R}^{q\times q}_{\mathrm{sym}} : M\mathbb{R}^q \subset \mathbb{V}\right\} \quad \text{for } \mathbb{V} \in \mathcal{V}_q. \end{aligned}$$

Necessary and sufficient condition for $Q \in \mathcal{Q}_\rho$:

Condition 0 (for Case 0) For any $\mathbb{V} \in \mathcal{V}_q$ with $1 \leq \dim(\mathbb{V}) < q$,

$$Q(\mathbb{M}(\mathbb{V})) \;<\; \frac{\dim(\mathbb{V})}{q}.$$

Condition 1 (for Case 1) For any $\mathbb{V} \in \mathcal{V}_q$ with $0 \leq \dim(\mathbb{V}) < q$,

$$Q(\mathbb{M}(\mathbb{V})) \;<\; \frac{\psi(\infty) - q + \dim(\mathbb{V})}{\psi(\infty)}.$$

References

Arslan, O., Constable, P. D. L. and Kent, J. T. (1995). Convergence behavior of the EM algorithm for the multivariate t-distributions. *Communications in Statistics – Theory and Methods* **24** 2981–3000. MR1364707

Arslan, O. and Kent, J. T. (1998). A note on the maximum likelihoood estimators for the location and scatter parameters of a multivariate Cauchy distribution. *Communications in Statistics – Theory and Methods* **27** 3007–3014. MR1659367

Auderset, C., Mazza, C. and Ruh, E. A. (2005). Angular Gaussian and Cauchy estimation. *J. Multivar. Anal.* **93** 180–197. MR2119770

Bhatia, R. (2007). *Positive definite matrices. Princeton Series in Applied Mathematics*. Princeton University Press, Princeton, NJ. MR2284176

Croux, C., Rousseeuw, P. J. and Hössjer, O. (1994). Generalized S-estimators. *J. Amer. Statist. Assoc.* **89** 1271–1281. MR1310221

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. With discussion. MR0501537

Dudley, R. M. (1968). Distances of probability measures and random variables. *Ann. Math. Statist* **39** 1563–1572. MR0230338

Dudley, R. M. (2002). *Real analysis and probability. Cambridge Studies in Advanced Mathematics* **74**. Cambridge University Press, Cambridge. Revised reprint of the 1989 original. MR1932358

Dudley, R. M., Sidenko, S. and Wang, Z. (2009). Differentiability of t-functionals of location and scatter. *Ann. Statist.* **37** 939–960. MR2502656

Dümbgen, L. (1998). On Tyler's M-functional of scatter in high dimension. *Ann. Inst. Statist. Math.* **50** 471–491. MR1664575

Dümbgen, L., Nordhausen, K. and Schuhmacher, H. (2013). New algorithms for M-estimation of multivariate scatter and location. *ArXiv Preprint, arXiv:1312.6489*.

Dümbgen, L. and Tyler, D. E. (2005). On the breakdown properties of some multivariate M-functionals. *Scand. J. Statist.* **32** 247–264. MR2188672

Eaton, M. L. (1989). *Group invariance applications in statistics. NSF-CBMS Regional Conference Series in Probability and Statistics* **1**. Institute of Mathematical Statistics, Hayward, CA; American Statistical Association, Alexandria, VA. MR1089423

Eriksen, P. S. (1987). Proportionality of covariance matrices. *Ann. Statist.* **15** 732–748. MR888437

Flury, B. K. (1986). Proportionality of k covariance matrices. *Statist. Probab. Lett.* **4** 29–33. MR822722

Haberman, S. J. (1989). Concavity and estimation. *Ann. Statist.* **17** 1631–1661. MR1026303

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust statistics. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*. John Wiley & Sons, Inc., New York. The approach based on influence functions. MR829458

Hettmansperger, T. P. and Randles, R. H. (2002). A practical affine equivariant multivariate median. *Biometrika* **89** 851–860. MR1946515

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statistics* **19** 293–325. MR0026294

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101. MR0161415

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1** 799–821. MR0356373

Huber, P. J. (1981). *Robust statistics. Wiley Series in Probability and Mathematical Statistics*. John Wiley & Sons, Inc., New York. MR606374

Jensen, S. T. and Johansen, S. (1987). Estimation of proportional covariances. *Statist. Probab. Lett.* **6** 83–85. MR907265

KENT, J. T. and TYLER, D. E. (1988). Maximum likelihood estimation for the wrapped Cauchy distribution. *J. Appl. Statist.* **15** 247–254.

KENT, J. T. and TYLER, D. E. (1991). Redescending M-estimates of multivariate location and scatter. *Ann. Statist.* **19** 2102–2119. MR1135166

KENT, J. T., TYLER, D. E. and VARDI, Y. (1994). A curious likelihood identity for the multivariate t-distribution. *Comm. Statist. Simulation Comput.* **23** 441–453. MR1279675

KOTZ, S. and NADARAJAH, S. (2004). *Multivariate t distributions and their applications*. Cambridge University Press, Cambridge. MR2038227

LANGE, K. L., LITTLE, R. J. A. and TAYLOR, J. M. G. (1989). Robust statistical modeling using the t distribution. *J. Amer. Statist. Assoc.* **84** 881–896. MR1134486

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of point estimation*, second ed. *Springer Texts in Statistics*. Springer-Verlag, New York. MR1639875

MARONNA, R. A. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Statist.* **4** 51–67. MR0388656

NIEMIRO, W. (1992). Asymptotics for M-estimators defined by convex minimization. *Ann. Statist.* **20** 1514–1533. MR1186263

NORDHAUSEN, K., OJA, H. and OLLILA, E. (2008). Robust independent component analysis based on two scatter matrices. *Australian J. Statist.* **37** 91–100.

PAINDAVEINE, D. (2008). A canonical definition of shape. *Statist. Probab. Lett.* **78** 2240–2247. MR2458033

SERFLING, R. J. (1980). *Approximation theorems of mathematical statistics*. *Wiley Series in Probability and Mathematical Statistics*. John Wiley & Sons, Inc., New York. MR595165

SIRKIÄ, S., TASKINEN, S. and OJA, H. (2007). Symmetrised M-estimators of multivariate scatter. *J. Multivariate Anal.* **98** 1611–1629. MR2370110

SKOROHOD, A. V. (1956). Limit theorems for stochastic processes. *Teor. Veroyatnost. i Primenen.* **1** 289–319. MR0084897

TYLER, D. E. (1987a). A distribution-free M-estimator of multivariate scatter. *Ann. Statist.* **15** 234–251. MR885734

TYLER, D. E. (1987b). Statistical analysis for the angular central Gaussian distribution on the sphere. *Biometrika* **74** 579–589. MR909362

TYLER, D. E., CRITCHLEY, F., DÜMBGEN, L. and OJA, H. (2009). Invariant co-ordinate selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 549–592. MR2749907

VAN DER VAART, A. W. (1998). *Asymptotic statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge University Press, Cambridge. MR1652247

WATSON, G. S. (1983). *Statistics on spheres*. *University of Arkansas Lecture Notes in the Mathematical Sciences* **6**. John Wiley & Sons, Inc., New York. A Wiley-Interscience Publication. MR709262

WIESEL, A. (2012). Geodesic convexity and covariance estimation. *IEEE Trans. Signal Process.* **60** 6182–6189. MR3006411